DeuCze

Julius-Maximilians-Universität Würzburg

UNI WÜ

Slezská univerzita v Opavě

Band 2

Josef Molnár

Studien zur Aufbereitung und Auswertung von Korpustexten

www.deucze.org

ISSN: 2190-9555

DeuCze

Korpuslinguistik Deutsch-Tschechisch kontrastiv

Das DeuCze-Publikationsforum dient der Veröffentlichung von Arbeiten, die im Zusammenhang mit dem Projekt 'DeuCze. Korpuslinguistik Deutsch-Tschechisch kontrastiv' entstehen. In diesem Projekt kooperieren Sprachgermanisten/-innen der Schlesischen Universität Opava und der Julius-Maximilian-Universität Würzburg. Durch die elektronische Publikation können Forschungsergebnisse und Diskussionsbeiträge leicht und in kurzer Zeit die wissenschaftliche Öffentlichkeit erreichen und das fachliche Gespräch anregen.

Herausgeber:
Iva Kratochvílová, Veronika Kot lková und Gabriela Rykalová (Opava) sowie Norbert Richard Wolf, Werner Wegstein und Peter Stahl (Würzburg)

Zitation dieser Publikation:

Molnár, Josef (2010). Studien zur Aufbereitung und Auswertung von Korpustexten. Schriftenreihe DeuCze, Band 2. Würzburg: Universität Würzburg.
ISBN: 978-3-923959-76-1

Josef Molnár

Studien zur Aufbereitung und Auswertung von Korpustexten

Disertační práce

Inauguraldissertation zur Erlangung der Würde eines Dr. phil.

Slezská univerzita – Opava – Filozoficko-přírodovědecká fakulta v Opavě

Julius-Maximilians-Universität – Würzburg – Philosophische Fakultät I

2010

Školitel / Betreuer: Prof. Dr. Werner Wegstein

Oponenti /
Erstgutachter: Prof. Priv.-Doz. PhDr. Lenka Vaňková, Dr.
Zweitgutachter: Prof. Dr. Wolf Peter Klein

Obhajoba / Tag des Kolloquiums: 20. 9. 2010

Abstrakt

Die vorliegende Arbeit beschäftigt sich mit dem Sprachkorpus aus zwei Blickwinkeln. Im technischen Teil handelt es sich um die Aufbereitung der Texte für das deutsch-tschechische Korpus DeuCze. Es wird hier der Vorgang von der Digitalisierung der Bücher bis zum Erstellen der wohlgeformten und validen XML-Dateien beschrieben. Diese Dateien sind bis zur Satzebene segmentiert und ermöglichen auf diese Weise die parallele Anzeige der Texte der beiden verglichenen Sprachen nach einzelnen Segmenten. Im analytischen Teil wird die Aufmerksamkeit der sprachlichen Analyse des Phänomens der Themaentwicklung innerhalb eines ausgewählten Textes gewidmet. Das Ziel sind also sowohl die aufbereiteten Dateien für das genannte Korpus als auch die Analyse der Teilthemaentwicklung.

Schlüsselwörter: Analyse der Korpustexte, DeuCze-Korpus, Teilthemaentwicklung, Parallelkorpus, Aufbereitung der Texte für das Sprachkorpus, Text, Thema

Abstract

This dissertation examines a linguistic corpus from two points of view. The technical part deals with the processing of the texts for the German-Czech DeuCze corpus, and describes the way to process the texts from digitalisation to well-formed and valid XML files. The files are segmented onto the sentence level, due to which it is possible to display the corpus text of the two languages segment-by-segment. In the analytical part the attention is focused on the linguistic analysis of the phenomenon of the part-theme development inside the selected text. The outcome of the dissertation are the processed files of the above mentioned corpus as well as the analysis of the part-theme development in the text.

Keywords: analysis of corpus texts, DeuCze corpus, progression of the part-theme, parallel corpus, processing of corpus texts, text, text theme

An dieser Stelle möchte ich mich für die Betreuung dieser Arbeit, freundliches Entgegenkommen und geduldige Unterstützung beim Schreiben dieser Dissertation bei Herrn Prof. Dr. Werner Wegstein bedanken.

Ein besonderer Dank gilt Frau Priv.-Doz. PhDr. Iva Kratochvílová, Ph.D. und Herrn Professor Prof. em. Dr. Dr. h.c. mult. Norbert Richard Wolf für ihr Interesse am Fortgang der Arbeit und wertvolle Anregungen.

Der Gemeinnützigen Hermann-Niermann-Stiftung danke ich für die finanzielle Unterstützung meines Stipendiumaufenthalts an der Julius-Maximilians-Universität in Würzburg, an der ich ein Jahr des Studiums verbringen konnte.

Allen Kollegen, die mich durch ihre Gespräche über das Studium und die Arbeit aufgemuntert haben, und denen, die mir durch das aufmerksame und schnelle Korrekturlesen meines Manuskriptes geholfen haben, danke ich herzlich.

Vielen Dank an alle Freunde, die sich für meine Arbeit interessierten und mich durch ihre Freundschaft unterstützten.

Ich danke meinem Vater und meiner ganzen Familie für ihre Unterstützung bei meinem Studium.

Inhalt

Tabellenverzeichnis

Verzeichnis der Abkürzungen

B/W	Black and White
COSMAS	Corpus Search, Management and Analysis System
CSS	Cascading Style Sheet
ČNK	Český národní korpus
DeReKo	Deutsches Referenzkorpus
DFG	Deutsche Forschungsgemeinschaft
dpi	Dots per inch
DTP	Desktoppublishing
F	Stellvertretende Bezeichnung für die Datei `KR_plain_to_xml_funktionen.php`
FG	Kürzt den Beleg *Friedhofsgesellschaft* ab
GIMP	GNU Image Manipulation Program (Softwarename)
GGU	Abkürzung für das Buch ‚Günter Grass: *Unkenrufe*' als Quellangabe bei den Belegen
IDS	Institut für Deutsche Sprache
JPG	Bildformat – ‚Joint Photographic Experts Group'
KB	Kilobyte
MB	Megabyte
MS	Microsoft (Firmenname)
OCR	Optical character recognition
PDF	Portable Document Format
PHP	Bezeichnung einer Skriptsprache
ppi	Pixel per inch
PTX	Stellvertretende Bezeichnung für die Datei `KR_plain_to_xml_kr_de.php`
TEI	Text Encoding Initiative
TIFF	Bildformat ‚Tagged Image File Format'
UK	Unmittelbare Konstituente
UTF	Unicode Transformation Format
XML	Extensible Markup Language
XSLT	XSL Transformation (XSL = Extensible Stylesheet Language)
XYQ	Zufällige Zeichenfolge in der Funktion einer Hilfsmarkierung

< >	Spitze Klammern kennzeichnen die XML-Tags
[…]	Auslassung bei den zitierten Belegen
Σ	Summe

Anhangsverzeichnis

Einleitung

Die Korpuslinguistik geht von der empirischen Untersuchung des Sprachmaterials aus. Bei den sprachwissenschaftlichen Untersuchungen wird also das beobachtet, was tatsächlich im Sprachgebrauch bereits verwendet wurde; mehrere Informationsquellen über dieses Gebiet der Linguistik sind bei Biber (2005), Lemnitzer / Zinsmeister (2006), McEnery (2003) oder Scherer (2006) zu finden. Diese Methodik ist heute deswegen so attraktiv, weil sie durch moderne Technologien neue Möglichkeiten bei der Sprachuntersuchung erschließt. Dies betrifft vor allem die Menge und Geschwindigkeit bei der Bearbeitung und Auswertung der zu untersuchenden Sprachdaten.

Die vorliegende Arbeit gliedert sich in zwei Teile. Im technischen Teil werden die Texte für das DeuCze-Korpus aufbereitet. Im analytischen Teil wird auf die Analyse der Teilthemaentwicklung eingegangen.

Der technische Teil hat das Ziel, die Charakteristik eines Sprachkorpus im Hinblick auf das deutsch-tschechische DeuCze-Korpus zusammenzufassen. Dieses Korpus ist in der Zusammenarbeit der germanistischen Abteilung der Universität Opava und des Lehrstuhls für deutsche Sprachwissenschaft der Universität Würzburg entstanden. Das Thema der Korpuslinguistik wird im ersten Kapitel vorgestellt, um diese Methode der Sprachuntersuchung einzuleiten. Neben einer kurzen Vorstellung der Korpuslinguistik richtet sich das Augenmerk auf das Sprachkorpus, das als Hilfe bei der Sprachanalyse verwendet wird. Es werden Kriterien vorgestellt, die ein Korpus als ein sprachwissenschaftliches Instrument erfüllen sollte. Aus der breiten Skala der unterschiedlichen Korpus-Arten werden ausgewählte Vertreter präsentiert, und das DeuCze-Korpus wird im Vergleich zu anderen Korpora vorgestellt, besonders im Rahmen der Korpusgröße und aus der Ebene der Annotation.

Das zweite Kapitel ist der Vorbereitung der Texte für das DeuCze-Korpus gewidmet. Als Primärquellen der Korpustexte werden belletristische Werke verwendet, die durch das Scannen digitalisiert und durch das Verfahren der optischen Zeichenerkennung in editierbare Textdateien umgewandelt werden. Vor der eigentlichen Digitalisierung der Bücher werden die Scaneinstellungen getestet, um optimale Werte für möglichst gute Ergebnisse bei der Bearbeitung zu erzielen. Die Grundlagen der

Bildbearbeitung stützen sich auf Autoren und Publikationen wie Kenney (2000), Waldraff (2004), Wegstein (2009) oder DFG-Praxisregeln Digitalisierung (2009).

Die Quelltexte für das Korpus werden als Textdateien gespeichert, deren Inhalt mit den Mitteln der Auszeichnungssprache XML (‚eXtensible Markup Language') kodiert wird. Die Transformation der Textdatei in eine XML-Datei wird im dritten Kapitel behandelt. Die Dateien bestehen einerseits aus dem Header (Dateikopf), der überwiegend die bibliografischen Angaben zu dem Buch enthält, andererseits aus dem eigenen Text. Der Textteil wird mit der Segmentierung der Texte bis zur Satzebene (Dubinin 2005) und mit Markierungen der typografischen Textgestaltung (Layout) ergänzt. Durch die Segmentierung ist es möglich, die Korpustexte, jeweils den Ausgangstext und seine Übersetzung, synoptisiert satzweise nebeneinander abzubilden.

Im vierten Kapitel werden die theoretische Ausgangsbasis und die Mittel für die Textanalyse dargestellt. Die Grundlage der Analyse bildet die deutsche Fassung des im DeuCze-Korpus enthaltenen Romans ‚Unkenrufe' von Günter Grass. Innerhalb des Romans, als eine abgeschlossene Texteinheit, wird das Thema analysiert. Jeder Text enthält ein Textthema als die größtmögliche Kurzfassung des Textinhalts. In der Regel wird das Textthema von Teilthemen begleitet (Brinker 2005). Es wird auch die Frage der Differenzierung des Textthemas und seine Aufteilung nach untergeordneten Elementen behandelt. Die Untersuchung des Teilthemas im Gesamttext wird im nächsten Kapitel durchgeführt.

Als Erstes müssen die relevanten Referenzbereiche in dem zu untersuchenden Text identifiziert werden. Zu diesem Zweck wird aus dem Gesamttext eine Frequenz-Wortformenliste gebildet, in der die Einträge nach ihrer Vorkommensfrequenz sortiert werden. Die am häufigsten vorkommenden Belege zeigen auf die für die Untersuchung relevanten Referenzbereiche.

Die Wortformenliste liefert aber nur grobe Ergebnisse, da sie Angaben ohne Kontext enthält. Damit die konkreten Belege und ihre Zusammenhänge im Text untersucht werden können, müssen auch die Beziehungen zwischen den Wörtern in Betracht gezogen werden. Die angeführten Arten der Wortschatzgruppierung erschließen die für die Analyse relevanten Wortzusammenhänge.

Die Beziehungen zwischen den einzelnen Wörtern im Text haben nicht nur die syntaktische Funktion, Sätze zu bilden. Diese Beziehungen helfen, durch ihre text-

konstituierende Funktion die Sätze zu einem Text zu verbinden. Bei dem Textverständnis und der Textkonstitution wird auf Wolf (2008) und Brinker (2005) zurückgegriffen. Diese Eigenschaften der Wörter werden aus der Sicht der Ausdrucksseite (Kohäsion) und auch der Inhaltsseite (Kohärenz) betrachtet.

Die Verbindungen der Wörter auf dem Gebiet der Kohärenz bilden Isotopien im Text. Die Isotopie wird in dieser Arbeit nach dem Verständnis von Greimas (1974) als die semantische Äquivalenz zwischen mindestens zwei sprachlichen Einheiten behandelt. Die Isotopiebeziehungen werden bei der Textuntersuchung besonders hilfreich, die konkreten Teilthemahinweise im Text zu zeigen. In einem selbstständigen Unterabschnitt werden die Wortbildungsprodukte als Bausteine des Textes behandelt. Die Wortbildung wird als Unterstützung bei der Identifizierung der Elemente eines Referenzbereiches, eines Teilthemas verwendet.

Im fünften Kapitel wird die Aufmerksamkeit dem Text- und Teilthema und ihrer Untersuchung gewidmet. Es werden die Erkenntnisse angewendet, die bei der Auseinandersetzung mit der Wortformenliste und den Arten der Wortschatzgruppierung gewonnen wurden. Bei dem Textthema, das sich auf den ganzen Text bezieht, kann eine Aufteilung in untergeordnete Teilthemen beobachtet werden. Die oben genannten Mittel dienen dazu, das Teilthema im Textverlauf und seine Einführung, Beibehaltung, Entwicklung und seinen Abschluss im Textverlauf zu untersuchen (Hausendorf / Kesselheim 2008). Es wird beobachtet, ob und wie auch die Teilthemen geteilt werden können. Einleitend werden die unterschiedlichen Themahinweise vorgestellt und nachfolgend werden die einzelnen Phasen eines Teilthemas im Text beobachtet. Die Referenzen werden besonders unter dem Gesichtspunkt der textbezogenen Aktivierung semantischer Merkmale (Semrekurrenz) und der Referenzidentität untersucht. Im zweiten Abschnitt wird die Analyse eines Phänomens, das nur in einem Teil des Textes vorkommt, durchgeführt. Dabei soll die Vorgehensweise für die Untersuchung im Gesamttext entworfen werden. Im abschließenden Unterkapitel werden die konkreten Themahinweise, die sich auf ein Teilthema beziehen, im Gesamttext untersucht. Das Ziel der Untersuchung ist festzustellen, wie die konkreten Themahinweise realisiert werden.

1 Das Korpus in der Sprachwissenschaft

Diese Arbeit konzentriert sich auf die wissenschaftliche Beschäftigung mit Texten unter Verwendung der Mittel, die die modernen computerlesbaren Sprachkorpora anbieten. Ein Korpus bedeutet nicht das etwaige Zentrum des Interesses eines Sprachwissenschaftlers, es ist ein Instrument, ein Medium, das die gestellten linguistischen Fragen beantworten hilft.

Einleitend wird kurz auf die Korpuslinguistik und auf das Sprachkorpus eingegangen. Das Sprachkorpus wird nicht nur allgemein angeführt, sondern es werden auch Korpora nach ausgewählten Kriterien vorgestellt. Besonders wird auch das deutsch-tschechische Parallelkorpus DeuCze behandelt, für das die hier bearbeiteten Texte aufbereitet werden.

1.1 Korpuslinguistik

In der Sprachwissenschaft werden zwei grundlegende Vorgehensweisen unterschieden, nämlich der empirische und der rationalistische Weg (vgl. Lemnitzer / Zinsmeister 2006, 5ff). Die rationalistische Richtung lehnt sich an die theoretischen Grundlagen der Erforschung des Sprachvermögens an, das allen Menschen gemeinsam ist. Die Ergebnisse solcher Forschung können präskriptiv die Sprachverwendung begleiten, was sich besonders beim Zweitsprachenerwerb als nützlich erweisen kann, unter anderem können die Theorien durch das Überprüfen bestimmter Sprachbelege bestätigt (oder widerlegt) werden.

Die empirische Richtung konzentriert sich dagegen auf das Beobachten der realisierten Sprachäußerungen und ihre Beschreibung:

> *Empirical data enable the linguist to make statements which are objective and based on language as it really is rather than statements which are subjective and based upon the individual's own internalised cognitive perception of the language.* (McEnery / Wilson ^{2}2001, 103)

Diese Richtung ist deskriptiv und ihre Ergebnisse können dazu beitragen, dass Tendenzen in der Sprachentwicklung beschrieben werden können. Auf diese Weise werden die Grammatikregeln für einen sog. richtigen Sprachgebrauch dem konkreten sprachgeschichtlichen Stand angepasst. Aus der Charakteristik der beiden Vorgehensweisen resultiert, dass sich diese zwei Richtungen ergänzen können.

Der rationalistische Ansatz richtet sich also überwiegend auf das Sprachsystem als Gegenstand der Untersuchung, und beim empirischen Ansatz (die Herangehensweise der Korpuslinguistik) wird der Sprachgebrauch beschrieben (Scherer 2006, 3). Der linguistische Bereich, in dem die Korpora den Zugang zu dem untersuchten Sprachmaterial vermitteln, wird Korpuslinguistik genannt. Bei der Korpuslinguistik handelt es sich nicht um ein Gebiet der Linguistik, wie beispielsweise bei Syntax oder Semantik. Korpuslinguistik ist eine Methodologie, die in verschiedenen linguistischen Bereichen angewandt werden kann (vgl. McEnery / Wilson ^{2}2001, 2). Die Korpuslinguistik ist also „*die Beschreibung von Äußerungen natürlicher Sprachen, ihrer Elemente und Strukturen, und die darauf aufbauende Theoriebildung auf der Grundlage von Analysen authentischer Texte, die in Korpora zusammengefasst sind.* […] *Korpusbasierte Sprachbeschreibung kann verschiedenen Zwecken dienen, zum Beispiel dem Sprachunterricht, der Sprachdokumentation, der Lexikographie oder der maschinellen Sprachverarbeitung.*“ (Lemnitzer / Zinsmeister 2006, 9)

Die in einem Korpus enthaltenen Texte müssen bestimmten technischen Anforderungen entsprechen; was hiermit gemeint ist, wird an ausgewählten Korpora illustriert.

1.2 Das Sprachkorpus

Ein Sprachkorpus enthält real beobachtbare und verifizierbare Daten und wird als ein sprachwissenschaftliches Instrument bei der empirischen Untersuchung der Sprache bzw. ihrer Phänomene verwendet. Im Grunde genommen kann jede beliebige Sammlung von Texten als ein Sprachkorpus dienen, und die Basis empirischer (Sprach-)Untersuchungen darstellen. Aber weil das Korpus heutzutage eine spezifische Bedeutung gewinnt, wird es nach bestimmten Kriterien definiert.

> „*Ein Korpus ist eine Sammlung schriftlicher oder gesprochener Äußerungen in einer oder mehreren Sprachen. Die Daten des Korpus sind digitalisiert, d.h. auf Rechnern gespeichert und maschinenlesbar. Die Bestandteile des Korpus, die Texte oder Äußerungsfolgen, bestehen aus den Daten selbst sowie möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind.*“ (Lemnitzer / Zinsmeister 2006, 40)

Die Art und Anzahl der Kriterien kann je nach Verwendungszweck variieren. Eine digitalisierte Form ist zwar nicht unbedingt nötig, aber die elektronisch verwalteten Korpora bieten weit mehr Möglichkeiten im Rahmen der Suche im Text als die

Papierkorpora. Genannt sei nicht nur der Umfang der in einem digitalen Korpus enthaltenen Daten, sondern auch die Geschwindigkeit und Präzision bei der Sortierung und Auswertung der Suchergebnisse. Der Computer ermöglicht es, nach einzelnen Wörtern, Wortgruppen oder (falls annotiert) nach den jeweiligen Wortarten oder Satzgliedern zu suchen. Dabei kann eine statistische Auswertung der Suchanfrage, wie die Anzahl der gesuchten, im Korpus enthaltenen Belege, besonders schnell berechnet werden. Ein weiterer bedeutender Vorteil der maschinellen Verarbeitung ist, dass die Daten schnell sortiert werden (z. B. nach dem Alphabet oder nach der Frequenz). Oder dass eine Anzeige ohne oder mit dem unterschiedlich breiten Kontext realisiert wird (mehr dazu McEnery / Wilson ^{2}2001, 18).

Gewöhnlich werden die Texte mindestens mit Metainformationen versehen, wie bibliografische Angaben und strukturelle Markierung (‚mark-up') mit einer Auszeichnungssprache, wie hier XML (‚eXtensible Markup Language'), um die Segmentierung der Quelldaten nach typografischen Kriterien (z. B. Kennzeichnung der Kapitel, Absätze und Sätze) einzutragen. Die Korpusdaten können mit linguistischer Annotation versehen werden, die die unterschiedlichen sprachlichen Merkmale näher bestimmt, wie z. B. die grammatischen Angaben bei den einzelnen Wörtern. Die unannotierten Texte werden bei den Analysen auch verwendet, aber die Möglichkeiten, die Daten nach bestimmten Merkmalen (z. B. nach Wortarten) zu untersuchen, sind kleiner. In diesem Projekt werden die Korpustexte bis zur Satzebene segmentiert, damit im Korpus eine parallele Anzeige der zwei bearbeiteten Sprachen nach einzelnen Sätzen realisiert werden kann. Bevor es auf die Kodierung der Korpustexte eingegangen wird, werden noch die Korpustexte und Korpusarten angesprochen.

1.2.1 Allgemeine Charakteristik der Korpustexte

Es hängt von dem jeweiligen Untersuchungszweck ab, welche Texte in ein Korpus aufgenommen werden (wie z. B. Nachrichten aus Zeitungen oder auch belletristische Werke). Alle zu untersuchenden Daten werden als Primärdaten bezeichnet (vgl. Scherer 2006, 3). Zu den Primärdaten treten zusätzliche Angaben hinzu, einerseits sind es die Metadaten, andererseits können auch andere (grammatische) Annotationen verwendet werden. Die Metadaten geben solche Informationen an, die bei der Entstehung eines Korpus von Bedeutung sind, wie z. B. Quellenangaben

zu den Texten, Arbeitsschritte, die im Header des jeweiligen Dokuments gespeichert werden, aber auch die strukturelle Beschreibung der Texte mit der metasprachlichen Markierung (vgl. Lemnitzer / Zinsmeister 2006, 44ff).

Die im Text implizit enthaltene Information kann durch die Annotation explizit gemacht werden (vgl. McEnery / Wilson ^{2}2001, 32ff). Es ist erstrebenswert, dass die Annotation der Texte bestimmten Standards folgt, weil auf diese Weise bearbeiteter Text den interessierten Benutzern leichter zugänglich sein kann. Die Metadaten und die grammatische Annotation können in den Korpustexten unterschiedlich kodiert werden. In diesem Projekt wird von der Markierung (auch ‚mark-up' oder Tagging genannt) mit XML ausgegangen. Die XML-Kodierung bietet einen internationalen Standard für die maschinell lesbaren Texte an. Eine konkrete Anwendung dieser Technologie bietet in ihren Empfehlungen ‚The Text Encoding Initiative' (TEI). Das TEI-Konsortium entwickelt die Kodierungsstandards für die digitale Textdarstellung in Bereichen der Geisteswissenschaften, Gesellschaftswissenschaften und Linguistik[1]. Das Ziel von TEI ist es, den Austausch der elektronischen Texte mit Standards zu unterstützen (vgl. McEnery / Xiao / Tono 2006, 23ff).

Das TEI-Format für die Kodierung der Korpustexte stellt nicht die einzige der Möglichkeiten dar, die auf diesem Gebiet zur Verfügung stehen. Je nach den Bedürfnissen der Wissenschaftler im jeweiligen Projekt können unterschiedliche Ebenen und Details des Inhalts und der Form der Annotation entsprechend kodiert werden. Es seien hier zwei weitere Standards, die innerhalb der EU ihre Verwendung finden, genannt. Von der Europäischen Union wurde die Initiative EAGLES (‚Expert Advisory Groups on Language Engineering Standards')[2] berufen, die Empfehlungen für Tagging (besonders im Bereich der Wortarten-Annotation) ausgibt. Ein anderer Standard, der sich auf die Kodierung der sprachlichen Korpora konzentriert, ist CES (‚Corpus Encoding Stanrard')[3]. CES wird auf (Sprach-)Korpora innerhalb der EU angewandt, dabei wird sowohl die Struktur, d. h. die Festlegung des formalen Aufbaus eines Korpus (Character Sets, Header Tags etc.), als auch die Kodierung der sprachlichen Infor-

1 Vgl.: http://www.tei-c.org/index.xml, zit.: 13. 7. 2009.
2 Mehr zu EAGLES unter http://www.ilc.cnr.it/EAGLES/home.html, zit.: 29. 12. 2010.
3 Mehr zu CES unter http://www.cs.vassar.edu/CES, zit.: 29. 12. 2010.

mationen definiert. Weil CES von TEI ausgeht, ist es also mit TEI kompatibel (vgl. McEnery / Wilson 2001, 34ff).

In diesem Projekt wird für das DeuCze-Korpus die Kodierung (bzw. die XML-Markierung der Quelldaten) nach den Regeln von TEI verwendet. Dies geschieht nicht nur aufgrund der Erfahrung der Würzburger Projektteilnehmer, sondern auch, weil TEI den Bedürfnissen des Projekts (wie Segmentierung der Texte bis zur Satzebene und Beschreibung der Elemente durch angeführte Attribute) entspricht und auch weitere Verarbeitung (wie Bearbeitung der Texte mit XSLT) erlaubt.

Bei den XML-Dokumenten wird standardmäßig die Zeichenkodierung Unicode UTF-8[4] verwendet, weil es Buchstaben und andere Zeichen einer großen Anzahl der Weltsprachen abbilden kann (vgl. McEnery / Xiao / Tono 2006, 27f). Auch für das Korpus DeuCze wird dieser Zeichensatz verwendet.

Neben der elektronischen Aufbereitung der Texte werden auch andere Kriterien vorgestellt, nach denen ausgewählte Korpusarten zu unterscheiden sind. Eine Klassifikation der Korpuskriterien führt Scherer an: „*Wichtig für die Konzeption eines Korpus sind Größe und Inhalt des Korpus sowie dessen Beständigkeit und Repräsentativität.*“ (Scherer 2006, 5) Es hängt von dem jeweiligen Zweck des bestimmten Korpus ab, welche Markierungen und Annotation eingesetzt werden. Noch vor Beginn der gegebenen Text-Untersuchungen werden Ziele gesetzt und erst nach Erstellung der Kriterien wird geplant, welche Metadaten eingetragen und welche konkreten Texte gewählt werden. Auf der anderen Seite können Informationen fehlen, die zusätzlich nicht mehr zu gewinnen sind und deswegen ist es günstiger, eher mehr Metadaten zu implementieren. Im DeuCze-Korpus wird in dieser Bearbeitungsphase keine grammatische Annotation durchgeführt, es wird aber geplant, in späteren Projektphasen die einzelnen Wörter nach Wortarten zu annotieren.

Für die linguistische Untersuchung ist es interessant, wenn das gegebene Korpus Texte von mehreren Autoren enthält. Das DeuCze-Korpus beinhaltet zurzeit Texte von vier Autoren.

Ein Korpus kann nicht eine gesamte Sprache repräsentieren, schon aufgrund der Tatsache, dass ständig neue Äußerungen gebildet werden. Der Grad der Reprä-

4 Unicode ermöglicht ‚nicht-romanische‘ und andere Buchstaben und Zeichen so darzustellen, wie sie in ihrer Schriftform sind (vgl. McEnery / Wilson [2]2001, 44). Die Nummer 8 gibt die Anzahl der Bits an, die für die Kodierung eines Zeichens verwendet werden.

sentativität hängt eng mit dem Verwendungszweck zusammen. Das ganze Werk eines Autors genügt zwar nicht dafür, die Sprache einer Zeitperiode adäquat zu vertreten, es kann aber ausreichend bei einer Untersuchung des Wortschatzes des gegebenen Autors sein (vgl. McEnery / Wilson ^{2}2001, 29f). Die Größe oder der Umfang des Korpus ist ein Kriterium, das je nach dem Verwendungszweck des Korpus seine Geltung ändern kann. Eine universelle Angabe des Umfangs der gesammelten Daten ist also nicht möglich. Zu beachten ist die Tatsache, dass ein Phänomen nur in einem ausreichenden Kontext eindeutig identifiziert werden kann. Ein Korpus ist also erst dann groß genug, wenn das untersuchte Phänomen auch oft genug vorkommt, um seine Regularität zu beobachten (vgl. Lemnitzer / Zinsmeinster 2006, 41).

Das DeuCze-Korpus ist dem Umfang nach ein Kleinkorpus, es wird auch aus der Perspektive der Annotation mit ausgewählten Korpora verglichen.

1.2.2 Korpusarten

Die zusammenfassende Charakteristik eines Korpus ändert sich je nach dem Verwendungszweck. In diesem Unterkapitel werden nur ausgewählte Korpusarten aus der gesamten Vielfalt der Sprachkorpora näher vorgestellt. Die Kriterien, nach denen die einzelnen Arten der Korpora unterschieden werden können, sind z. B. das Speichermedium (computerlesbares vs. nicht computerlesbares Korpus), die Aufbereitung der Quelldaten (annotiertes vs. nicht annotiertes Korpus), das Sprachmedium (Korpus der geschriebenen Sprache vs. Korpus der gesprochenen Sprache), der zeitliche Bezug (Korpus der Gegenwartssprache vs. historisches Korpus) oder die Anzahl der Sprachen (ein- vs. mehrsprachiges Korpus) (vgl. Scherer 2006, 16ff).

Bei mehrsprachigen Korpora wird zwischen Parallelkorpora (engl. ‚parallel corpora‘)[5] und vergleichbaren Korpora (engl. ‚translation corpora‘)[6] unterschieden. Die Parallelkorpora enthalten Texte in Ausgangssprache und ihre Übersetzungen in der Zielsprache. Es können ausgewählte Einheiten der beiden Sprachen einander zugeordnet und synoptisiert werden, d. h. eine parallele (synoptisierte) Anzeige der Sätze oder

5 Auf Deutsch Parallelkorpora (vgl. Scherer 2006, 29f) bzw. Übersetzungskorpora (vgl. Kratochvílová 2006, 35ff).

6 In anderen Quellen auch ‚Comparable corpora‘ genannt (vgl. McEnery 2003, 450). Auf Deutsch vergleichbare Korpora (vgl. Scherer 2006, 29f), Vergleichskorpora (vgl. Lemnitzer / Zinsmeister 2006, 104) bzw. Kontrastkorpora (vgl. Kratochvílová 2006, 35ff).

Wörter ist möglich. Die vergleichbaren Korpora enthalten nicht die Übersetzung, sondern sie enthalten Texte des gleichen Genres bzw. der gleichen Struktur. Bei der Textuntersuchung im Bereich der vergleichbaren Korpora sind die Ergebnisse nicht von den sprachlichen Fähigkeiten des Übersetzers beeinflusst, weil beide (u. U. auch mehrere) bearbeitete Texte in der Ausgangssprache stehen (vgl. KRATOCHVÍLOVÁ 2006, 35ff).

In dieser Arbeit werden nur ausgewählte Korpora vorgestellt, die Auswahlkriterien sind der Umfang der enthaltenen Texte und der Grad der Annotation, damit das DeuCze-Korpus im Kontrast zu den angeführten Korpora vorgestellt werden kann. Nach dem Umfang der enthaltenen Daten wird zwischen Groß- und Kleinkorpora unterschieden. Die großen Korpora bieten heutzutage mehrere Millionen an enthaltenen Wörtern. Trotzdem entstehen kleinere, spezialisierte Korpora, die für bestimmte Untersuchungszwecke erstellt werden, um den Ansprüchen der konkreten Fragestellung besser zu entsprechen (vgl. MCENERY / WILSON ^{2}2001, 189ff). Als Beispiel für Großkorpora werden das Tschechische Nationalkorpus und die Korpora des Instituts für Deutsche Sprache in Mannheim stellvertretend angeführt. Im Bereich der unterschiedlichen Stufen der Annotation der Korpustexte werden die Korpora TIGER-Korpus und Europarl angeführt.[7] Das in diesem Projekt bearbeitete DeuCze-Korpus gehört zu den Kleinkorpora und die Segmentierung seiner Texte reicht bis zur Satzebene.

Das Tschechische Nationalkorpus

Český národní korpus (ČNK) ist ein akademisches Projekt, das überwiegend tschechisch geschriebene Texte enthält. Es wird von dem Institut des Tschechischen Nationalkorpus an der Philosophischen Fakultät der Karlsuniversität in Prag (ÚČNK) bearbeitet. Das Gesamt-Korpus besteht aus mehreren Korpora:

a) geschrieben; synchron: SYN2009PUB, SYN2006PUB, SYN2005, SYN2000, FSC2000, KSK-DOPISY, ORWELL – im Durchschnitt fast jeweils 200 Millionen Wörter, die meisten Texte stammen aus dem Zeitraum von 1990 bis einschl. 2004.

7 Eine umfangreichere Übersicht von deutschen Einzelkorpora findet sich z. B. bei LEMNITZER / ZINSMEISTER 2006, 115–126.

b) gesprochen; synchron: ORAL2008, ORAL2006, PMK, BMK – im Durchschnitt fast jeweils 800 Tausend Wörter

c) diachrones Korpus: DIAKORP – 1,6 Millionen Wörter

d) paralleles Korpus: InterCorp – 44 Millionen Wörter[8]

Von den genannten Kategorien ist nur die Kategorie der geschriebenen synchronen Korpora größtenteils lemmatisiert und enthält morphologische Markierung. Das parallele Korpus InterCorp enthält diese Art der Metadaten nur teilweise und der Rest (gesprochen synchron, diachron) enthält keine Metadaten. Das Spektrum der vertretenen Texte ist breit und reicht von Publizistik, Belletristik, der privaten Korrespondenz bis hin zu den Sprachaufnahmen.

Die ausgewählten Korpora enthalten morphosyntaktische Markierungen[9]. Die entsprechenden Tags sind durch Attribute mit morphologischen Angaben ergänzt, die das Wort morphologisch identifizieren, und auch die lemmatisierten Formen sind eingefügt. Während der Analyse werden die Wörter ohne Kontext bearbeitet, deswegen können die Ergebnisse uneindeutig sein. Jedes Wort enthält also eine aus 16 Zeichen bestehende Marke, die die morphologischen Kategorien angibt. Eine detaillierte Übersicht zu dem System der Markierung steht im Internet auf der Webseite des Projektes[10] zur Verfügung.

Die Korpustexte sind über das Programm Bonito zugänglich. Die Suchmöglichkeiten bieten die Anfragen z. B. nach einzelnen Wörtern, Mehrwortverbindungen, Lemmata oder nach den morphologischen Markierungen. Die Suche mittels der regulären Ausdrücke ist auch verfügbar und kann eventuell auf ausgewählte Korpora begrenzt werden. Die Suchergebnisse können mit unterschiedlich langem Kontext angezeigt werden bzw. es können Kollokationen (links oder rechts) aufgesucht werden. Dies ist nur eine Auswahl der Suchmöglichkeiten, eine detaillierte Übersicht befindet sich im on-line erreichbaren Bonito-Handbuch[11].

Innerhalb des Tschechischen Nationalkorpus können auch ausgewählte Fremdsprachen untersucht werden. Das Projekt Interkorp stellt sich das Ziel, „ein umfangreiches paralleles synchrones Korpus, das möglichst viele Sprachen enthält“,

8 Detailliertere Angaben unter: http://ucnk.ff.cuni.cz/struktura.php, zit.: 7. 5. 2010.
9 Die Beschreibung der Markierungen: http://ucnk.ff.cuni.cz/bonito/znacky.php, zit.: 7. 5. 2010.
10 Siehe http://ucnk.ff.cuni.cz/bonito/znacky.php, zit.: 7. 5. 2010.
11 Siehe http://ucnk.ff.cuni.cz/bonito/index.php, zit.: 7. 5. 2010.

aufzubauen[12], und für die parallele Anzeige ist es über eine spezielle Schnittstelle „Park“ zugänglich. Das Korpus wird ständig ausgebaut und erweitert. Das Korpus ermöglicht die Suche nach Wortformen, Mehrwortverbindungen und Lemmata, bei ausgewählten Sprachen ist die Suche nach morphosyntaktischen Tags möglich, außerdem können reguläre Ausdrücke verwendet werden[13].

Korpora des Instituts für Deutsche Sprache (IDS) in Mannheim

Die Korpora der geschriebenen Sprache sind als das Deutsche Referenzkorpus – DeReKo zugänglich; insgesamt sind über 3,9 Milliarden Wörter enthalten, die vertretenen Textsorten werden v. a. durch belletristische, wissenschaftliche und populär-wissenschaftliche Texte repräsentiert[14].

Das im Korpus der gesprochenen Sprache enthaltene Material wird in zwei Ausgangsgruppen geteilt: Sprachvarietäten und Gesprächskorpora. Die vertretenen Sprachvarietäten verteilen sich weiter auf binnendeutsche Mundarten, innerdeutsche Umgangssprachen / Standardsprache, auslandsdeutsche Varietäten und Sonstiges (wie z. B. Slawische Mundarten im Ruhrgebiet). Die Gesprächskorpora bieten Aufnahmen aus den Bereichen wie Dialogstrukturen, Beratungsgespräche oder Kindersprache[15].

Das historische Textkorpus enthält Texte aus dem Zeitraum von 1700 bis etwa 1918 und umfasst 45 Millionen Wörter[16]. Aus urheberrechtlichen Gründen ist im Internet zur Recherche durch Cosmas II nur ein Teil dieses Korpus zugänglich.

Die Recherche erfolgt mittels der (im IDS entwickelten) Software COSMAS II. Dieses Programm setzt virtuelle Korpora zusammen, die dem Benutzer zu der konkreten Recherche dienen. Diese Software kann auf Windowsrechnern installiert werden, kann aber auch in der webbasierten Version[17] verwendet werden, die über das Fenster des gängigen Webbrowsers aktiviert werden kann.

Die Korpora sind lemmatisiert und morphologisch analysiert. Diese Tatsache erweitert die Suchmöglichkeiten, weil neben der einfachen Stichwortsuche auch nach

12 Vgl. http://korpus.cz/intercorp-info.php, zit.: 7. 5. 2010.
13 Eine detailliertere Beschreibung auf den Seiten des Projektes: http://korpus.cz/intercorp-info.php, zit.: 7. 5. 2010.
14 Vgl. http://www.ids-mannheim.de, zit.: 7. 5. 2010.
15 Eine detailliertere Übersicht siehe unter: http://agd.ids-mannheim.de/html/korpora/korpus-index.shtml zit.: 7. 5. 2010.
16 Vgl.: http://www.ids-mannheim.de/lexik/HistorischesKorpus, zit.: 7. 5. 2010.
17 Der Zugang ist von dieser Adresse möglich: https://cosmas2.ids-mannheim.de/cosmas2-web

Wortformen, nach Lexemen oder auch nach einzelnen Wortbildungsaffixen gesucht werden kann. Es bietet auch die Möglichkeit, Platzhalter zu verwenden, die die ausgewählten Buchstaben oder Wortteile ersetzen. Darüber hinaus ist es auch möglich, nach Wortabfolgen oder Wortkombinationen zu suchen (vgl. Scherer 2006, 80ff).

TIGER-Korpus

Das Korpus TIGER Treebank enthält sprachwissenschaftlich annotierte Texte[18]. In der Version 2.1 sind 50 000 Sätze aus der Zeitung Frankfurter Rundschau enthalten. Die Texte wurden halb automatisch mit Informationen zu den Wortarten und zur syntaktischen Struktur, nach den Regeln der Dependenzgrammatik, getaggt. Daneben sind auch morphologische und Lemma-Angaben zu den Terminalknoten enthalten.

In diesem Korpus wird mittels des Programms TIGER-Search recherchiert, das auf dem Rechner installiert werden muss. Diese Software ermittelt dann die Sätze, die das untersuchte Phänomen enthalten.[19] Jeder Satz wird in der Form von einer Baumstruktur abgebildet, so werden die syntaktischen Beziehungen visualisiert.

Die Abfrage realisiert sich durch Eingabe der Wortarten und ihrer unterschiedlichen Beziehungen im Satz. Die Suchabfragen werden mit eigener Abfragesprache formuliert, nähere Regeln sind im TIGERSearch Manual zu finden.[20]

Europarl

Das Korpus European Parliament Proceedings Parallel Corpus 1996–2009, das sog. ‚Europarl', enthält die Sitzungsberichte der Europäischen Union. Es enthält Versionen in 11 europäischen Sprachen. Das Ziel dieser Bearbeitung ist es, ein System zu erstellen, das als Textbasis für maschinelle Übersetzung dient. Für diesen Zweck wurden entsprechende Texteinheiten – die Sätze – extrahiert und mit der Auszeichnungssprache XML markiert.[21] Diese alignierten Texte eignen sich außer zur Übersetzung auch zu sprachwissenschaftlichen Untersuchungen im Bereich des Sprachvergleichs der vertretenen Sprachen.

18 Vgl.: http://www.ims.uni-stuttgart.de/projekte/TIGER, zit.: 18. 11. 2010.
19 Vgl.: http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/oldindex.shtml, zit.: 18. 11. 2010.
20 TIGERSearch Manual in HTML Version: http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/manual_html.html, zit.: 18. 11. 2010.
21 Siehe: http://www.statmt.org/europarl, zit.: 18. 11. 2010.

Nicht alle enthaltenen Sitzungsberichte sind in allen vertretenen Sprachen zugänglich, deswegen unterscheidet sich die Größe der jeweiligen Versionen. Als Beispiel wird hier die englische und deutsche Version genannt, das Tschechische ist hier nicht vertreten. Die englische Version enthält 1 891 918 und die deutsche Version 1 822 735 Sätze. Die Sprachkombination Deutsch-Englisch enthält dagegen nur 1 581 107 Einträge, weil nur für diese Anzahl der Sätze die jeweilige Übersetzung enthalten ist. Auf der Website des Projektes stehen die Texte zur Verfügung, sie können mit unterschiedlichen Programmen bearbeitet werden, wie z. B. Satzaligner oder eigene Programme zur Suche im Text.[22]

Die Texte wurden halb automatisch nach Sätzen segmentiert, der Ausgangspunkt bei der Bestimmung der Satzgrenze ist der Punkt, es muss dabei zwischen Satzendepunkt und Punkt bei einer Abkürzung unterschieden werden. Anhand der Satzlänge werden die Sätze in den unterschiedlichen Sprachen automatisch einander entsprechend zugeordnet.[23]

DeuCze

Das DeuCze-Korpus ist ein Vertreter der Kleinkorpora (engl. ‚small corpora'; vgl. Wolf 2010, 22f). Es entstand im Rahmen eines Projektes, das in der Zusammenarbeit der germanistischen Abteilung der Universität Opava und des Lehrstuhls für deutsche Sprachwissenschaft der Universität Würzburg durchgeführt wird. Innerhalb dieses Projektes wurde ein synchrones, deutsch-tschechisches Parallelkorpus der geschriebenen Texte erstellt. Das Korpus enthält belletristische Texte, zurzeit sind es zwei deutsche Romane mit ihren Übersetzungen ins Tschechische und zwei tschechische Romane mit ihren Übersetzungen ins Deutsche und es wird künftig erweitert. Im Rahmen dieser Arbeit wurden folgende Werke technisch bearbeitet:

- Jiří Kratochvil: *Nesmrtelný příběh aneb Život Soni Trocké-Sammlerové čili Román karneval.* Verlag Petrov, Brno 2005 und die Übersetzung ins Deutsche von Liedtke, Kathrin und Vagadayová, Milka: *Unsterbliche Geschichte oder Das Leben von Sonja Trotzkij-Sammler oder Karneval.* Verlag Amman, Zürich 2000.

22 Siehe: http://www.statmt.org/europarl, zit.: 18. 11. 2010.
23 Näheres zu dem Arbeitsverfahren in Koehn 2005, 2f.

- Thomas Brussig: *Am kürzeren Ende der Sonnenallee.* Fischer Taschenbuch Verlag, Frankfurt am Main 2001 und die Übersetzung ins Tschechische von Zoubková, Jana: *Na kratším konci ulice.* Verlag Odeon, Praha 2001.

Alle vier bearbeiteten Texte stehen digitalisiert zur Verfügung, sie sind in Form von XML-Dateien und bis zur Satzebene segmentiert. Das Korpus ist mittels eines gängigen Web-Browsers im Internet unter der Adresse ‚deucze.org' zugänglich. Das Bestreben im Bereich des Umfangs der Kleinkorpora besteht nicht darin, die Gesamtheit der Sprache zu repräsentieren, sondern nur einen bestimmten Teil der Sprache zu erfassen. Die größten Vorteile der kleinen Korpora bestehen darin, dass sie als projektgebundene Spezialkorpora erstellt werden und daher den Anforderungen des jeweiligen Untersuchungsziels besser entsprechen können. Auf diese Weise können ganze Strukturen und Konstruktionen kontext- und textgebunden untersucht werden (s. Kratochvílová 2010, 172ff). Für zeit- oder raumlinguistische Analysen sind kleine Korpora ausreichend. Aber bei morphosyntaktischen Analysen muss häufig auf ein großes Korpus zurückgegriffen werden, beispielsweise, wenn bei der Analyse an einem kleinen Korpus vorläufige Ergebnisse ermittelt werden, die bei einem großen Korpus erweitert untersucht werden (vgl. Wolf 2010, 22ff).

Das DeuCze-Korpus erschließt die enthaltenen Texte online in einem Browserfenster für die sprachwissenschaftliche Untersuchung[24]. Die automatisch geregelten Suchmöglichkeiten können eingeschränkt sein. Die Begrenzung besteht darin, dass in vielen Fällen die Anfragen an den Text mit einem Suchprogramm durchgeführt werden, und dieses sucht im Text entweder nach den Zeichenketten oder nach bestimmten Informationen, die als Annotation dem Quelltext hinzugefügt wurden. Die Markierung und Annotation wird bereits vor dem Korpusbau geplant und kann eventuell im Laufe eines Projektes angepasst und erweitert werden. In dieser Projektphase wird in den Texten für das DeuCze-Korpus nur die Markierung der Struktur bis zur Satzebene durchgeführt. Diese Segmentierung dient der Synoptisierung der Korpustexte bei der zweisprachigen Anzeige. Eine grammatische Annotation ist zurzeit nicht enthalten, und deswegen ist die Suche nach Wortarten nicht möglich.

24 Das Korpus ist unter der Internetadresse ‚deucze.org' zugänglich.

1.3 Zusammenfassung

Dieses Kapitel stellte das Sprachkorpus vor. Im ersten Unterkapitel wurde die Position des Korpus in der Sprachwissenschaft bzw. in der Korpuslinguistik behandelt. Im zweiten Unterkapitel wurde das Sprachkorpus nicht nur allgemein behandelt, sondern auch konkrete ausgewählte Korpora wurden näher vorgestellt.

Besondere Aufmerksamkeit wurde der Unterscheidung zwischen den Groß- und Kleinkorpora und dem Niveau der Annotation der Korpustexte gewidmet. Auch wenn die Großkorpora mehrere Millionen Wörter enthalten, haben die Kleinkorpora eigene Vorteile. Die kleinen Spezialkorpora können nämlich besser der spezifischen sprachwissenschaftlichen Fragestellung dienen. Ähnlich ist es bei der Annotation der Korpustexte, denn nur entsprechend annotierte Korpora können spezifische Fragen beantworten.

Das in diesem Projekt bearbeitete DeuCze-Korpus ist ein Parallelkorpus, es enthält Texte im Tschechischen und im Deutschen. Die Texte sind in DeuCze im XML-Format gespeichert und für die Anzeige bis zur Satzebene segmentiert. Zurzeit ist keine grammatische Annotation enthalten, und ihre Ergänzung ist im Rahmen dieser Arbeit nicht mehr vorgesehen. Eine detailliertere Annotation der Wörter z. B. nach Wortarten wird für eine spätere Projektphase geplant. Die Transformation der Texte, von dem gedruckten Buch bis in die XML-Datei spielt sich in mehreren größeren Schritten ab.

2 Die Vorbereitung der Texte für das DeuCze-Korpus

Die Quelltexte für das Korpus stehen als gedruckte Bücher zur Verfügung und müssen zunächst digitalisiert werden. Zu Anfang wird der Schritt behandelt, wie die Texte jeweils in eine maschinenlesbare Textdatei umgewandelt werden. Das Ziel der Gestaltung der später im Korpus enthaltenen Daten ist, dass die Elemente im kodierten Text dem Layout in der Originalvorlage entsprechen. Die im Folgenden behandelten Hauptaspekte sind das Scannen, die optische Zeichenerkennung und die weitere elektronische Textverarbeitung.[25] Zuerst werden die Quelltexte vorgestellt.

2.1 Beschreibung der Primärquellen

Im Rahmen dieser Arbeit werden zwei belletristische Werke jeweils mit ihren Übersetzungen (also insgesamt vier Bücher) bearbeitet:

a) tschechischer Roman mit deutscher Übersetzung:

Jiří Kratochvil: *Nesmrtelný příběh aneb Život Soni Trocké-Sammlerové čili Román karneval*. Verlag Petrov, Brno, 2005 – zweite Auflage[26] (im Verlag Petrov die erste Auflage), S. 11–213. Der Verlag stellte diese Publikation als eine PDF-Datei zur Verfügung.

Jiří Kratochvil: *Unsterbliche Geschichte oder Das Leben der Sonja Trotzkij-Sammler oder Karneval*. Übersetzt von Liedtke, Kathrin und Vagadayová, Milka. Erste Auflage, Zürich 2000 (die Übersetzung geht von der Originalausgabe aus dem Jahre 1997 – Verlag Atlantis, Brno aus), S. 7–296.

Der Roman ist in fünf Teile (Bücher) geteilt, und insgesamt sind in diesem Werk 66 Kapitel enthalten. Die Romanbücher und die einzelnen Kapitel sind mit Überschriften versehen, der Gesamttext ist nach Absätzen geteilt. Im Text kommen zentriert gesetzte Gedichte vor. Es ist auch ein Brief vertreten, der Sondergestaltung, z. B. bei der Grußformel am Abschluss, verlangt. In diesem Text sind Stellen enthalten, die in Kursiv-, Kapitälchen- oder Fettschrift formatiert sind.

---

25 Die Beschreibung dieser Projektphase lehnt sich in manchen Punkten an die Ausführungen in „Digitalisierung von Primärquellen für die TextGrid-Umgebung: Modellfall Campe-Wörterbuch“ an. Zugänglich unter WWW: http://www.textgrid.de/fileadmin/TextGrid/veroeffentlichungen/Leitlinien-zur-Kodierung-des-Campe-Woerterbuchs-in-TEI-P5.pdf

26 Aus technischen Gründen war es nicht möglich, die erste Auflage innerhalb dieses Projektes zu bearbeiten.

b) deutscher Roman mit tschechischer Übersetzung:

Thomas Brussig: *Am kürzeren Ende der Sonnenallee.* Fischer Taschenbuch Verlag, Frankfurt am Main, 2001, S. 7–157.

Thomas Brussig: *Na kratším konci ulice.* Übersetzt von Zoubková, Jana. Verlag Odeon, Praha, 2001, S. 9–103.

Dieser Roman ist in 14 Kapitel geteilt, und jedes Kapitel beginnt mit einer Überschrift. Der Text wird nach Absätzen geteilt, Textpassagen in Kursivschrift und ein zentriert gesetztes Gedicht sind auch vertreten.

Das Layout des jeweiligen Textes wird im Header der konkreten XML-Datei beschrieben. Die Informationen zur Textgestaltung der jeweiligen Bücher sind für Planung der XML-Struktur nützlich.

Die spätere Beschreibung der Arbeitsschritte geht von dem deutschen Text von Kratochvil aus. Im Anhang Nr. 1 werden Beispielseiten aus diesem Buch angeführt. Es können dabei folgende Phänomene betrachtet werden: Überschrift des Buches als Romanteil, Anfang eines Kapitels bzw. auch der erste Absatz innerhalb eines Kapitels, ein Gedicht und der Brief.

Die zu erstellenden Bilddateien müssen bestimmten Kriterien entsprechen, damit die Korpustexte mit möglichst kleinem Aufwand weiterverarbeitet werden können.

2.2 Ziele der Digitalisierung

Bei den gescannten Bildern handelt es sich um das Konzept der Pixelgrafik (vgl. Waldraff 2004, 3). Die Pixel oder Bildpunkte sind die Minimaleinheiten, die die kleinsten darstellbaren Details eines Bildes vertreten. Erst durch Verwendung entsprechend vieler Bildpunkte können die enthaltenen Linien oder Rundungen kontinuierlicher abgebildet werden, was auch bei der OCR-Bearbeitung (englisch: OCR – ‚optical character recognition' / optische Zeichenerkennung) eine Rolle spielt. Die Pixelanzahl wird durch Auflösung ausgedrückt (s. auch Brunner / Ecker 2001, 72).

Die behandelten Bücher befinden sich in einem guten Zustand, das Papier hat keine Zusätze wie z. B. Flecken. Um den Text für die weitere Bearbeitung zu erhalten, werden die für das DeuCze-Korpus bearbeiteten Bücher seitenweise gescannt, indem jede Buchseite als ein eigenständiges Bild gespeichert wird.

Im Allgemeinen können mit der Scansoftware bestimmte Scaneinstellungen entsprechend angepasst werden. In den nächsten Abschnitten werden die Merkmale angesprochen, die für die Zwecke der geplanten OCR-Bearbeitung und Bildschirmpräsentation der Bilder von Bedeutung sind.

Das Ziel beim Scannen ist es, qualitativ hochwertige Bilder, die ‚digital master' (vgl. Kenney 2000, 25), für die Archivzwecke und weitere Bearbeitung zu gewinnen. Die konkreten Scaneinstellungen wie Auflösung oder Speicherformat werden durch Probescans bestimmt. Die Masterdateien bilden die Grundlage für die weitere Bearbeitung und werden nicht mehr geändert. Zu dem Zweck der weiteren Verwendung werden kleinere Kopien, sogenannte Derivative, gebildet[27], die als Ausgangspunkt für die grafische Darstellung der Bücher im Rahmen der Benutzerschnittstelle des Korpus dienen. Bei kleineren Kopien muss dann über das Netz keine so große Datenmenge mehr übertragen werden, es verkürzt sich also die Übertragungs- und Darstellungszeit der Bilder. Es handelt sich um möglichst kleine Bild-Dateien mit angemessen lesbarer Textdarstellung. Die weitere Bearbeitung der Bilder besteht darin, dass sie mit OCR-Software gelesen werden, um den Text der Romane in der Form von editierbaren Textdateien als Grundlage für das DeuCze-Korpus zu erhalten. Um die Archiv-Dateien nicht zu beschädigen, wird auch bei der OCR-Bearbeitung nur mit Kopien gearbeitet. Die optimalen Eigenschaften dieser Typen der Bilddateien werden im selbstständigen Abschnitt über Probescans behandelt, vorher richtet sich die Aufmerksamkeit noch auf die nötigen Schritte bei der Digitalisierung der gedruckten Texte.

2.3 Bilddigitalisierung

In diesem Projekt werden die ausgewählten gedruckten Texte durch das Scannen digitalisiert, die folgenden Ausführungen beschäftigen sich mit allgemeinen Voraussetzungen für die Bilddigitalisierung. Diese Tätigkeit wurde von Hilfskräften der Germanistikabteilungen in Opava und in Würzburg durchgeführt. Das Ziel der Digitalisierung ist es, Bilder in entsprechender Qualität für Archivzwecke und weitere Bearbeitung zu gewinnen, um bei der späteren Behandlung möglichst gute Ergebnisse zu erreichen. Für das Scannen wurde der Scanner mit Buchkante ‚Plustek OpticBook

27 Vgl. auch DFG-Praxisregeln „Digitalisierung", S. 7. Zugänglich unter WWW: http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/download/praxisregeln_digitalisierung.pdf

3600‘[28] verwendet – ein zum Scannen von Büchern spezialisierter Flachbettscanner, außerdem der Scanner ‚HP tma 8200‘. Die verwendete Scansoftware ist ‚Book Pilot‘, bei dem HP-Scanner die HP-Scansoftware. Die geeigneten Farbtiefen- und Auflösungswerte und das Speicherformat der Bilddateien werden durch Probescans getestet. Vor dem eigentlichen Scannen müssen bestimmte Vorüberlegungen zur Digitalisierung der Bücher behandelt werden, d. h. vor allem die Fragen, wie sich die Farben und die Auflösung auf das Scanresultat auswirken.

2.3.1 Farbeinstellungen

Bereits beim Scannen muss an die folgende Verwendung der gewonnenen Bilder gedacht werden. In späteren Bearbeitungsschritten werden die Bilddateien (die Seiten der Bücher) mit einer OCR-Software gelesen, um den Text zu extrahieren. Das OCR-Ergebnis hängt mit der Anzahl der auf dem Bild enthaltenen Farben, der Farbtiefe, zusammen. Mehrere Farbstufen bedeuten eine höhere Farbtiefe, und auf diese Weise können auch kleinere Details im Bild entsprechend dargestellt oder gar mit dem verwendeten OCR-Programm korrekt entschlüsselt werden.

Die Farbinformation wird in der Datei durch Bits eingeschrieben. Die Bits, die die kleinsten Informationseinheiten darstellen, können durch die Werte des Binärsystems ‚0‘ und ‚1‘ zwei Farben repräsentieren, d. h., auf diese Weise wird eine bitonale Grafik – mit Schwarz und Weiß – kodiert. Bei den elektronischen Bildern werden Farbabstufungen erzielt, indem ein Pixel durch mehrere Bits repräsentiert wird. Es wird mit Abstufungen der zwei Grundfarben gearbeitet und zu den Farben Schwarz und Weiß treten Dunkel- und Hellgrau hinzu (vgl. Waldraff 2004, 9ff). Das hier verwendete 8-Bit-Grau wird von 256 Grauabstufungen repräsentiert.

Das unten auf Bild Nr. 1 angeführte Beispiel zeigt einen (vergrößerten) Textausschnitt[29], der mit dem schwarz-weißen (links) und dem grauen (rechts) Scanmodus gescannt wurde:

28 Mehr zu dem Scanner auf der Internetadresse des Herstellers: http://www.plustek.de, zit.: 6. 9. 2009.

29 Der Beispieltext stammt aus Kratochvil, Jiří: *Unsterbliche Geschichte oder Das Leben der Sonja Trotzkij-Sammler oder Karneval.* Zürich 2000, S. 13 (Ausschnitt aus Zeilen 16 und 17).

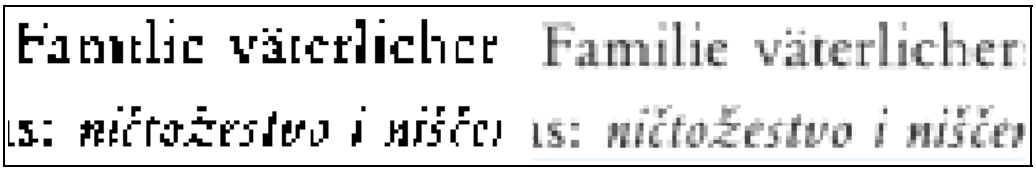


Bild 1: Auflösung 100 dpi

Niedrigerer Auflösungswert in Kombination mit wenigen Farbstufen kann zu Informationsverlust führen, was hier z. B. bei dem Hatschek, der im Textausschnitt links über „*c*“ und „*z*“ im Wort „*ničtožestvo*“ steht, zu sehen ist. Dagegen tragen mehrere Farbabstufungen auf dem Textausschnitt rechts zur besseren Lesbarkeit bei. An dieser Probe wird deutlich, dass die Bildqualität (und das OCR-Ergebnis) in Abhängigkeit von der gewählten Anzahl der Farben unmittelbar beeinflusst wird und dass der graue Scanmodus bessere Ergebnisse auch bei niedrigerer Auflösung erreicht.

Durch die Zuweisung eines Farbindex können die Bits Farbwerte erhalten. Bei den Texten, die in dieser Arbeit behandelt werden, kommen neben der schwarzen Druckfarbe keine weiteren Farben vor. Eine farbige Darstellung der Texte hat hier keine Verwendung und bleibt, bis auf Probescans, unbeachtet.

Grundsätzlich werden beim Scannen folgende Farbenmodi angeboten: Farbscan (‚Color Scan‘), Graustufenscan (‚Grayscale Scan‘) und Textscan (‚B/W Scan‘). Die in Klammern stehenden Bezeichnungen entstammen dem verwendeten Programm ‚Book Pilot‘ und können je nach Programm variieren. Mit den Probescans werden alle drei Farbenmodi getestet, um die möglichen Vor- und Nachteile zu beobachten. Neben der Farbtiefe ist auch die Auflösung bedeutend, weil diese Eigenschaft direkt die Bildqualität beeinflusst.

2.3.2 Auflösung

Ein gescanntes Bild besteht aus einer bestimmten Anzahl von Bildelementen, den Pixeln, die die kleinsten adressierbaren Elemente eines Bildes darstellen (vgl. Maschke 2004, 138). Durch eine höhere Anzahl an Bildelementen in einer Pixelgrafik können feinere Details aufgezeichnet werden. Von dieser Tatsache ausgehend, sei die Auflösung definiert „[…] *als die Zerlegung einer Informationsmenge in getrennt wahrnehmbare Elemente oder auch als das Vermögen, dicht an dicht liegende Objekte als eigenständig zu erfassen*“ (Waldraff 2004, 29).

Der Auflösungswert muss noch vor dem eigentlichen Scannen bestimmt werden, damit das Scan-Resultat auch die entsprechende Qualität für die gewünschte Weiterverarbeitung hat. Zu diesem Zweck werden die Probescans durchgeführt. Die Auflösung wird in Punkten pro Zoll bzw. als ‚dpi' (auf Englisch ‚dots per inch') angegeben. Schon die Bezeichnung dieser Einheit gibt an, dass sie ausdrückt, wie viele Bildelemente auf der Länge von einem Zoll (1 inch = 25,4 mm) verwendet werden (Kenney 2000, 32).

Die unterschiedlichen Scanner bieten gewöhnlich mehrere Stufen der Bit-Werte für die Kodierung der Farben, stellvertretend werden Optionen aufgeführt von 1 Bit (für schwarz-weiß), 8 Bit (für grau) oder 24 Bit (für Farbe). In Zusammenhang mit diesen Angaben ist auch die Speichergröße der Dateien zu erwähnen, eine kleine Übersicht ausgewählter Werte bietet die unten angeführte Tabelle (vgl. Waldraff 2004, 44):

Scanauflösung in dpi	Vorlagengröße in Megabyte (MB)		
	15 x 10 cm		
	1 Bit	8 Bit	24 Bit
100	0,03	0,26	0,77
300	0,3	2,3	6,9
600	1,2	9,3	27,8

Tabelle 1: Speichergröße im Vergleich

Die Tabelle zeigt die Werte der Speichergröße der Dateien in MB. Es handelt sich um eine Scanvorlage von 15 x 10 cm, die mit Auflösungswerten von 100, 300 und 600 dpi erfasst und im Format TIFF gespeichert sind. Dabei wird zwischen den Farbtiefewerten von 1, 8 und 24 Bit unterschieden. Im Vergleich zu den farbigen Scans haben die grauen Bilder für die Bedürfnisse dieses Projektes eine ausreichende Qualität, weswegen sie vor den farbigen Bildern bevorzugt werden, die unnötig viel Speicherplatz einnehmen.

Ein höherer Bit-Wert beeinflusst positiv die weitere Bildbearbeitung der Texte mit der OCR-Software. Die Bilddateien im Rahmen dieses Projektes repräsentieren eine Textabbildung, deswegen sollte die Scanauflösung so gewählt werden, dass möglichst gute OCR-Ergebnisse gewonnen werden können. Im Allgemeinen kann davon ausgegangen werden, dass bei Graustufen eine Mindestauflösung von 300 dpi ausreicht

und erst bei den bitonalen Scans 600 dpi erforderlich sind.[30] Mit den Probescans wird geprüft, wie sich die unterschiedlichen Einstellungen auf das OCR-Resultat auswirken. Bevor die digitalisierten Bilder weiter verwendet oder archiviert werden, muss noch geprüft werden, ob sie den Ansprüchen an die Qualität entsprechen.

2.3.3 Qualitätskontrolle

Einen untrennbaren Teil der Digitalisierung bildet die Kontrolle, die je nach der geplanten Verwendung der gescannten Bilder bestimmte Eigenschaften überprüft (vgl. Kenney 2000, 61ff).

Bei dem Scannen von Texten werden bestimmte Grundregeln verwendet (vgl. Kraus 1998, 62), wie z. B. die Tatsache, dass die Vorlage möglichst gerade eingescannt werden soll, oder dass durch das Vergrößern des Kontrastes zwischen Hintergrund und Schrift bessere OCR-Ergebnisse zu erzielen sind – in diesem Projekt wird der Kontrastwert um 15 bis 20 % erhöht (die gescannte Seite scheint auf den ersten Blick schwarz-weiß zu sein). Schatten, die z. B. dadurch entstehen, dass das Buch nicht ausreichend auf das Scannerglas gedrückt wurde, werden wenn möglich mit einem Bildverarbeitungsprogramm ausgebessert; wenn die betroffene Seite jedoch den gestellten Kriterien im Ganzen nicht entspricht, wird sie erneut abgetastet. Die Kontrolle wird bei den Masterdateien durchgeführt, die Derivative werden erst von den kontrollierten Dateien gebildet und sollten deswegen nicht mangelhaft sein.

Nach dem Scannen muss also jede Seite einzeln am Bildschirm angesehen werden und es wird kontrolliert, ob die Seiten vollständig gescannt wurden, ob die Zeilen gerade sind und ob der Text der Seite vollständig erfasst wurde.

Hiermit ist die Behandlung der allgemeinen Scan-Einstellungen abgeschlossen. Im Folgenden wird näher auf das Scannen der Bücher und die anschließende Behandlung der Bilddateien für das bearbeitete Projekt eingegangen.

2.4 Textscannen und OCR-Textbearbeitung

Beim Scannen der Texte müssen Kriterien und Methode des Textscannens unter Berücksichtigung der konkreten Texte in diesem Projekt beachtet werden. Die Aufmerksamkeit richtet sich im Folgenden auf die Scaneinstellungen, die in Bezug auf die

30 Vgl. DFG-Praxisregeln „Digitalisierung“, 2009, 7.

optische Zeichenerkennung betrachtet werden. Es wird auch auf die Probescans näher eingegangen, um den Scanvorgang für die Bedürfnisse dieses Projektes zu optimieren.

2.4.1 Die Scanner-Einstellungen

Die Texte in den hier bearbeiteten Büchern sind schwarz auf hellem Papier gedruckt. Der Hintergrund erscheint durch die Scaneinstellungen (wie Kontrast-wertanpassung) weiß. Beim Scannen von Textvorlagen ist ausschlaggebend, dass die OCR-Software die abgebildeten Buchstaben möglichst exakt auflösen kann. Bei dem Scanmodus ‚Schwarzweiß' spielt auch der Schwellenwert eine Rolle. Alle Farb-abstufungen der jeweiligen Bildbereiche, die sich unter diesem Wert befinden, werden als Schwarz interpretiert, und die helleren werden in Weiß umgewandelt. Das Ergebnis stellt eine aus schwarzen und weißen Pixeln bestehende Bitmap dar; bei niedrigerer Auflösung droht Informationsverlust (vgl. Bild Nr. 2). Bei der Verwendung von Scanmodus ‚Grau' kann der Widerdruck, d. h., der Text auf der Rückseite des Blattes, sichtbar werden, dies kann durch die Einstellung des Kontrastwertes und zusätzlich mit einer bildverarbeitenden Software (wie Photoshop oder GIMP) verbessert werden. Da die Bildanpassungen die Daten beeinflussen, führt Kenney zwei Arten dieser Anpassungen an (vgl. Kenney 2000, 51). Einerseits sind das die Anpassungen, die akzeptiert werden können, wie Veränderungen des Kontrasts oder kleine Farbwert- und Tonwertveränderungen. Andererseits sind es fragliche Bildanpassungen, wie Bild-schärfen oder softwaregesteuerte Farbwert- oder Tonwerterhöhung. Bei den Materialien im Rahmen dieses Projektes wurde bei dem gewählten Farbtiefescanmodus ‚Grau' (direkt beim Scannen) nur der Kontrastwert angepasst[31]. Der größere Kontrast zwischen dem Text und der Farbe des Papiers bringt bessere OCR-Ergebnisse.

Die vier folgenden Beispielgrafiken sollen anhand von vier unterschiedlichen Scaneinstellungen illustrieren, wie Auflösung und Farbtiefe die Lesbarkeit beeinflussen[32]. Die Bilder stellen einen Beispiel-Textausschnitt dar, wobei jeweils das linke Bild 100%-Darstellungsgröße und das rechte Bild eine 300%-Vergrößerung abbildet. Die ersten zwei Bilder behandeln Fälle mit der Auflösung 100 dpi und die letzten zwei mit der Auflösung 300 dpi, jeweils zuerst schwarz-weiß und dann in grau.

31 Zum Thema ‚Bildanpassung' siehe auch DFG-Praxisregeln „Digitalisierung", 2009. S. 10.

32 Der Beispieltext stammt aus Kratochvil, Jiří: *Unsterbliche Geschichte oder Das Leben der Sonja Trotzkij-Sammler oder Karneval.* Zürich 2000, S. 13 (Zeile 16 und 17).

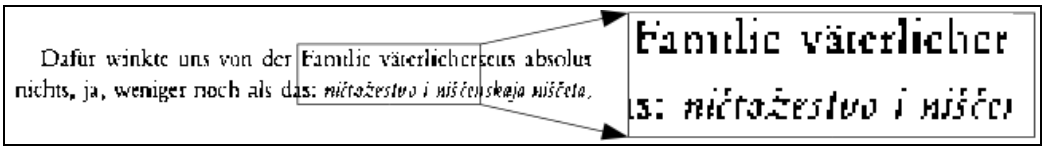


Bild 2: Gescannter Text, schwarz-weiß, Auflösung 100 dpi

- auf dem Bild Nr. 2 ist ein schwarz-weiß gescannter Text, die Auflösung ist 100 dpi – manche Details gehen an bestimmten Textstellen verloren, z. B. ist der Hatschek über dem „c“ und „z“ nicht eindeutig zu identifizieren.

Bild 3: Gescannter Text, grau, Auflösung 100 dpi

- das Bild Nr. 3 zeigt grau gescannten Text mit der Auflösung 100 dpi – dank mehreren Graustufen können beim Lesen auch kleinere Details unterschieden werden.

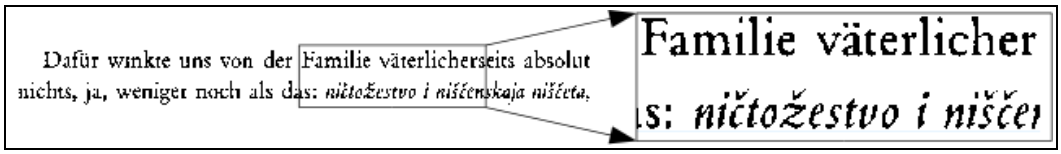


Bild 4: Gescannter Text, schwarz-weiß, Auflösung 300 dpi

- das Bild Nr. 4 stellt einen schwarz-weißen Text, die Auflösung beträgt 300 dpi – aufgrund der höheren Auflösung sind auch die Interpunktionszeichen lesbar erhalten.

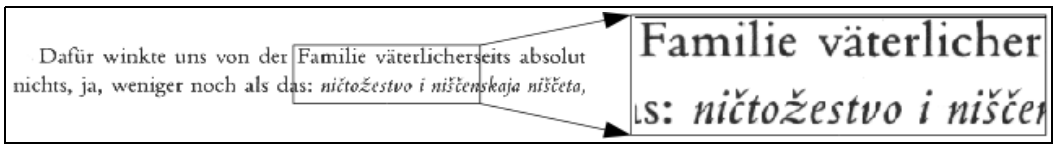


Bild 5: Gescannter Text, grau, Auflösung 300 dpi

- das Bild Nr. 5 zeigt einen grau gescannten Text; die Auflösung ist 300 dpi – die feinere Auflösung in Kombination mit Graustufen wirkt sich positiv auf eine gute Lesbarkeit des Textes aus, die Rundungen werden durch die Grauabstufungen geglättet und erscheinen relativ scharf.

Die Beispiele bestätigen die Tatsache, dass wenn im Schwarz-Weiß-Modus feinere Details dargestellt werden sollen, eine höhere Auflösung, z. B. auch 600 dpi, verwendet werden muss. Wenn der Grau-Modus gewählt wird, kann eine niedrigere Auflösung gewählt werden, denn bei einem Graustufenbild können die Ränder der

Schriftzeichen durch graue Pixel geglättet und schärfer abgebildet werden. Daraus resultiert, dass die größere Auflösung zur besseren Lesbarkeit beiträgt. Die angewendeten Scanner-Einstellungen haben also einen unmittelbaren Einfluss darauf, wie das Resultat der OCR-Textverarbeitung sein wird.

2.4.2 Das OCR-Verfahren

Das Ergebnis der optischen Texterkennung ermöglicht es, die gewonnenen Texte weiter als maschinell lesbare Textdateien zu speichern und als solche zu bearbeiten. Die gewählten Auflösungswerte müssen zu befriedigenden OCR-Ergebnissen führen. Eine wenn möglich endgültige Entscheidung für den konkreten Auflösungswert muss bereits vor dem Scannen getroffen werden; sollte dies nicht schon zu diesem Zeitpunkt bestimmt werden, ist es besser, eher eine höhere Scanauflösung zu wählen, die später verringert werden kann, umgekehrt geht es nicht.

Erfahrungsgemäß ist es nicht ausgeschlossen, bestimmte Auflösungswerte allgemein für die OCR-Bearbeitung vorzuschlagen, wie:

- 200 – 300 ppi[33] für Texte in Lesegröße 9–12 DTP-Punkte[34]
- 400 – 600 ppi für kleinere Schriften bis 8 DTP-Punkte

Die oben angeführten Werte (vgl. Waldraff 2004, 134) können als eine allgemeine Ausgangsbasis betrachtet werden, doch in einem konkreten Projekt ist es nützlich, Probescans durchzuführen. Ein untrennbarer Teil solcher Proben ist auch das OCR-Verfahren, um die optimalen Scaneinstellungen für die Transformation der Bilder in Text zu gewinnen. Die Qualität der Ergebnisse kann von verschiedenen Faktoren beeinflusst werden, nicht nur von der Schriftgröße und -art, sondern auch von Papier- oder Druckfarbequalität. Für die eigentliche OCR-Texterfassung werden die Masterdateien nicht verwendet, es wird nur mit abgeleiteten Kopien gearbeitet.

Das OCR-Verfahren spielt sich in mehreren Schritten ab, die der Benutzer nicht unbedingt als selbstständige Stufen wahrnimmt (vgl. Schlicht 1993, 72ff). Die Kenntnis dieser Stufen kann dazu beitragen, die Probleme, falls aufgetreten, gezielter zu beheben. Der erste Schritt bei der Texterfassung besteht darin, dass die Textvorlage mit einem Scanner eingelesen wird. Danach werden die Ergebnisse von einem speziellen Programm für optische Zeichenerkennung ausgewertet. Bei der Verarbeitung einer

33 ‚ppi': ‚pixel per inch', Bildpunkte pro Zoll.
34 DTP-Punkte: eine typografische Maßeinheit für Schriftgrade.

digitalen Textvorlage muss das verwendete Programm die für die OCR-Verarbeitung relevanten Objekte vom Hintergrund deutlich und eindeutig unterscheiden können. Den nächsten Schritt bildet die Differenzierung von wesentlichen (die zu erkennende Schrift) und unwesentlichen Bereichen (Schmutz, Grafiken). Das Programm geht so vor, dass es das eingescannte Bild nach bestimmten, vorgegebenen Mustern abtastet, d. h., die einzelnen Pixelansammlungen werden isoliert und dem gefundenen Muster werden entsprechende Zeichen zugeordnet (die eigentliche Texterkennung). Damit die Software in den segmentierten Bereichen einen Buchstaben erkennt, ist die Pixelmenge (die Auflösung) entscheidend. Eine Rolle spielt auch die verwendete Schriftart. Das verwendete Texterkennungsprogramm (ABBYY FineReader) hatte bei den hier bearbeiteten Texten keine wesentlichen Schwierigkeiten mit den vertretenen Standard-schriftarten. Eine weitere Eigenschaft der Dateien, die ihre Bearbeitung beeinflussen kann, sind die Speichermöglichkeiten der gewonnen Texte.

Speichern des Textes nach der OCR-Verarbeitung

Für das Speichern des Textes bietet FineReader mehrere Möglichkeiten an. Es wurde das MS Word Format (im FineReader als ‚DOC-Format' bezeichnet) gewählt, um die Formatierung zu erhalten. Weitere mögliche Speicher-Einstellungen für das Textformat sind ‚Genaue Kopie', ‚Bearbeitbare Kopie', ‚Formatierter Text' und ‚Nur Text'. Sie unterscheiden sich dadurch, wie viel Formatierung nach dem Speichern beibehalten werden soll. Im Rahmen der ‚Texteinstellungen' wird noch die Option ‚Seitenumbrüche beibehalten' gewählt. Das Ziel beim Speichern der Dokumente in diesem Schritt ist es, dass die Gestaltung der Seiten und Zeilen mit der Druckvorlage identisch bleibt und dass die Formatierung (Kursiv-, Fett-, Kapitälchenschrift) auch in die Ausgabedatei übernommen wird. Der Einzug der Absätze wurde nicht bei allen Absätzen konsequent erkannt und musste kontrolliert werden. (Dies wird später noch im Zusammenhang mit der Text-Korrektur behandelt). Es wurde die Speicher-Option ‚Formatierter Text' gewählt, und die fehlenden Informationen (z. B. Seitennummerierung oder Markierung von Absatzanfängen) wurden manuell eingefügt. Der gewonnene Text muss kontrolliert und falls nötig auch entsprechend angepasst werden.

Text-Korrektur

Um das OCR-Verfahren abzuschließen, müssen in den erstellten Texten alle eventuellen Abweichungen von dem Original korrigiert werden. Auch wenn die Ergebnisse oft zufriedenstellend sind, eine 100%ige Entsprechung wird durch das automatische Verfahren nicht erreicht. Das Layout muss entsprechend angepasst werden, und ebenso müssen alle Zeichen, die bei der optischen Zeichenerkennung nicht korrekt erkannt wurden, ausgebessert werden. Für die Fehlerkorrektur sind drei Schritte vorgesehen:

a) Das Programm ABBY FineReader vergleicht den erkannten Text mit dem eingebauten Wörterbuch. Es hebt die Wörter oder Zeichen hervor, die in seinem Wörterbuch nicht vertreten sind. Diese Hervorhebung signalisiert also die vermutlich nicht korrekt erkannten Stellen. Auf diese Weise werden die markantesten Fehler beseitigt.

b) Im zweiten Schritt wird der Text als eine Datei im Format des Textverarbeitungsprogramms MS Word gespeichert. Diese Software bietet die Möglichkeit, eine eingebaute Rechtschreibkorrektur zu nutzen. Durch die grafische Hervorhebung der vermeintlichen Fehler werden viele nicht korrekt erkannte Textstellen schnell aufgefunden. Es empfiehlt sich nicht, die Fehler automatisch korrigieren zu lassen, jede hervorgehobene Stelle sollte mit dem Original verglichen werden. Falls es im Original Satzfehler gibt, werden Korrekturen an dem gedruckten Text im TEI-Header der zum Schluss erstellten XML-Datei verzeichnet.

c) Das Korrekturlesen stellt den dritten Korrekturschritt dar. Auch wenn die Wörter der behandelten Dokumente in den Wortlisten der eingebauten Rechtschreibkorrektur von FineReader oder MS Word vertreten sind, gibt es Phänomene, die nicht gleich entdeckt werden. Es ist vor allem die mögliche falsche Verwendung von Wortformen (wie z. B. im Falle der Kongruenz von Subjekt und Prädikat) gemeint. Eine konkrete Abhilfe, die OCR-Fehler noch effektiver zu korrigieren, leistet das Korrekturlesen; diese Aufgabe, durch die sich die Fehlerwahrscheinlichkeit weiter verringert, wurde von den Hilfskräften durchgeführt.

In dieser Phase sollen auch die Worttrennung am Zeilenende und die Seitennummerierung beachtet werden. Wenn im Text leere Seiten ohne Nummerierung vorkommen, müssen die fehlenden Seitenzahlen manuell ergänzt werden. Dies betrifft die Fälle, wo in den bearbeiteten Texten am Ende eines Kapitels zwischen den

Romanteilen eine Leerseite vorkommt, damit das neue Kapitel auf der rechten Seite beginnt.

Beim Speichern der durch das komplexe OCR-Verfahren gewonnenen Texte muss gewährleistet werden, dass die Daten auch nach Langzeitspeicherung lesbar sind. Dies bedeutet, dass auch die Wahl der Zeichen-Kodierung (wie z. B. UTF-8) beachtet werden muss.

Speichern der korrigierten Textdatei

Nach den Korrekturschritten kann davon ausgegangen werden, dass der elektronische Text weitestgehend dem gedruckten Original entspricht. Für die zu speichernde Datei wurde der Zeichensatz Unicode UTF-8 gewählt, um zu gewährleisten, dass der Computer alle enthaltenen Zeichen der vertretenen Sprachen richtig interpretiert. Abschließend wird der Text in einer Textdatei ohne weitere zusätzliche Formatierung (engl. ‚plain text') gespeichert.

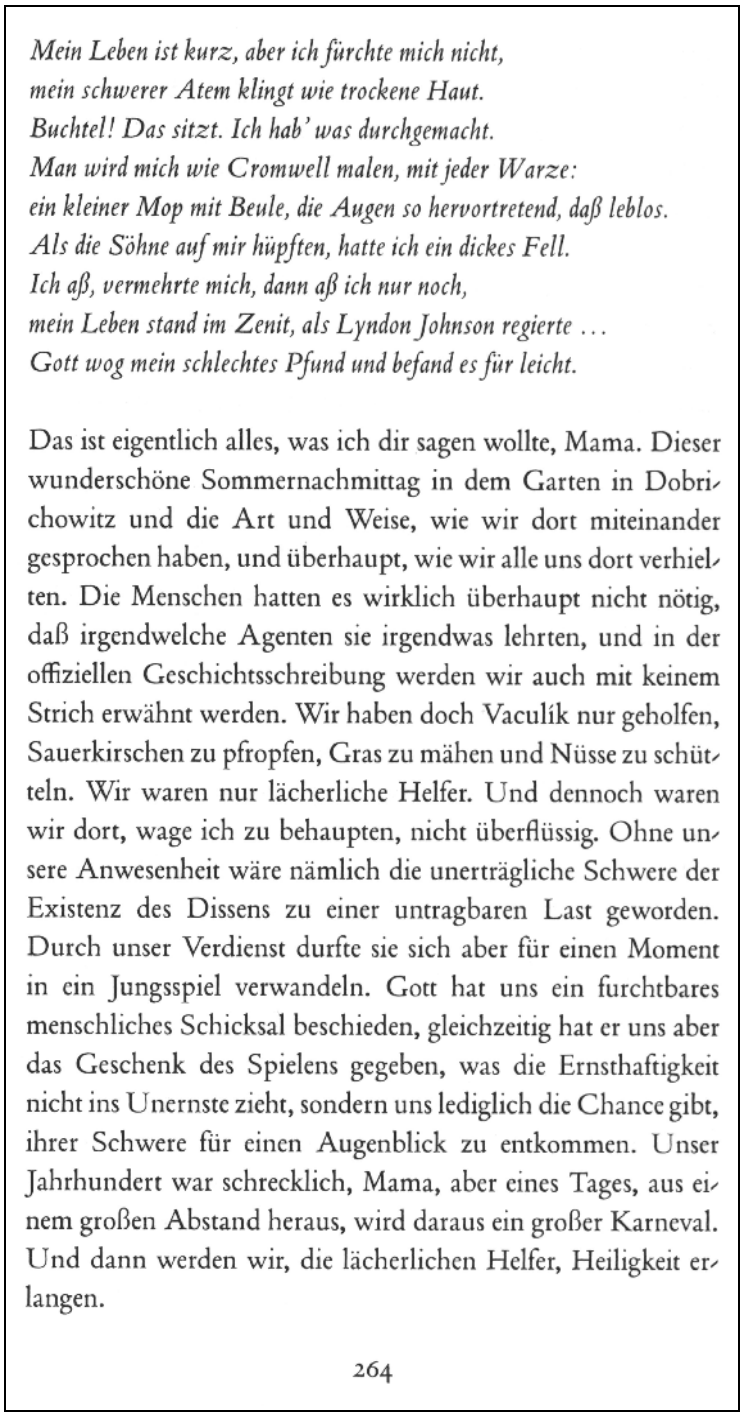
Durch die Verwendung des Datenformats ‚plain text' wird gewährleistet, dass in der Datei nur der Text ohne ergänzende Informationen (wie MS Word-Dokument-Header mit Formatierungsangaben) gespeichert wird. Dieses Format ist auch deswegen für die Dauerspeicherung geeignet, weil diese Art der Textdateien dann gängige Textverarbeitungsprogramme öffnen und bearbeiten können. Da eine solche Datei nicht an eine bestimmte Software oder ein bestimmtes Betriebssystem gebunden ist, ist es auch einfacher, sie mit anderen Projektteilnehmern zu teilen.

Zuletzt muss noch geprüft werden, wie sich die behandelten Einstellungen auf die weitere Verarbeitung der Texte auswirken.

2.4.3 Probescans

Durch die Probescans werden Auflösung, Kontrast und Format der zu speichernden Datei und ihr Einfluss auf das OCR-Ergebnis getestet. Die Analyse der Probescans soll Vorteile und Nachteile der zu verwendenden Einstellungswerte für das Scannen der Gesamttexte zeigen, um im Anschluss die Scaneinstellungen für die Masterdateien maximal zu optimieren. Die Grundlage für die Proben stellte die Seite264 (hier als Bild Nr. 6 dargestellt) aus dem Buch *Unsterbliche Geschichte* von Jiří Kratochvil dar, das als ein Bestandteil der Textbasis für das Korpus im Projekt DeuCze

dient. Diese Seite wurde deswegen gewählt, weil sie mehrere diverse Phänomene enthält, die bei der OCR-Bearbeitung möglicherweise Schwierigkeiten bereiten könnten. Besonders sind es die Satz- und andere Sonderzeichen (Komma, Punkt, Ausrufe- und Auslassungszeichen, Apostroph, Doppelpunkt, Trennstrich) sowie Normal- und Kursivschrift.



Mein Leben ist kurz, aber ich fürchte mich nicht,
mein schwerer Atem klingt wie trockene Haut.
Buchtel! Das sitzt. Ich hab' was durchgemacht.
Man wird mich wie Cromwell malen, mit jeder Warze:
ein kleiner Mop mit Beule, die Augen so hervortretend, daß leblos.
Als die Söhne auf mir hüpften, hatte ich ein dickes Fell.
Ich aß, vermehrte mich, dann aß ich nur noch,
mein Leben stand im Zenit, als Lyndon Johnson regierte …
Gott wog mein schlechtes Pfund und befand es für leicht.

Das ist eigentlich alles, was ich dir sagen wollte, Mama. Dieser wunderschöne Sommernachmittag in dem Garten in Dobrichowitz und die Art und Weise, wie wir dort miteinander gesprochen haben, und überhaupt, wie wir alle uns dort verhielten. Die Menschen hatten es wirklich überhaupt nicht nötig, daß irgendwelche Agenten sie irgendwas lehrten, und in der offiziellen Geschichtsschreibung werden wir auch mit keinem Strich erwähnt werden. Wir haben doch Vaculík nur geholfen, Sauerkirschen zu pfropfen, Gras zu mähen und Nüsse zu schütteln. Wir waren nur lächerliche Helfer. Und dennoch waren wir dort, wage ich zu behaupten, nicht überflüssig. Ohne unsere Anwesenheit wäre nämlich die unerträgliche Schwere der Existenz des Dissens zu einer untragbaren Last geworden. Durch unser Verdienst durfte sie sich aber für einen Moment in ein Jungsspiel verwandeln. Gott hat uns ein furchtbares menschliches Schicksal beschieden, gleichzeitig hat er uns aber das Geschenk des Spielens gegeben, was die Ernsthaftigkeit nicht ins Unernste zieht, sondern uns lediglich die Chance gibt, ihrer Schwere für einen Augenblick zu entkommen. Unser Jahrhundert war schrecklich, Mama, aber eines Tages, aus einem großen Abstand heraus, wird daraus ein großer Karneval. Und dann werden wir, die lächerlichen Helfer, Heiligkeit erlangen.

264

Bild 6: Probescan

Es wurde hier die deutsche Fassung behandelt, die Erkennungssprache bei dem verwendeten OCR-Programm ‚ABBY FineReader' wurde deswegen auf Deutsch eingestellt. Die getestete Textseite wurde mit dem Scanner ‚Plustek OpticBook 3600' und der Scansoftware ‚Book Pilot' mit unterschiedlichen Scan- und Speichereinstellungen abgetastet und schließlich mit ‚ABBY FineReader' gelesen. Im Folgenden werden die Anzahl der durch das OCR-Verfahren nicht korrekt erkannten Zeichen und das Speicherformat betrachtet, um die am besten geeignete Einstellung für die Masterdateien zu bestimmen. Durch die Proben wurden zwei Speicherformate getestet:

a) Das Format TIFF – es ist ein weitverbreitetes Standardformat. TIFF wird von vielen Anwendungen unterstützt und erlaubt die Bilder in unterschiedlichen Farbtiefen (bis zu 24 Bit) zu speichern; dabei kann gewählt werden, ob das Bild (verlustlos) komprimiert wird oder ob es ohne Kompression gespeichert wird (ausführlicher Osterberg 2005, 81). Dieses Format ist für die Masterdateien (Archivdateien) geeignet; Kenney führt an, dass das Masterdateienformat „*open and well-documented, widely supported, and cross-platform compatible*" (s. Kenney 2000, 51) sein sollte.

b) Das Format JPG – es verwendet verlustbehaftete Kompression, die die Dateigröße reduziert, ohne den Gesamteindruck des Bildes zu beeinträchtigen. Die beim Speichern verlorene Information ist nicht mehr herzustellen (vgl. Kraus 1998, 48). Dank des kleineren Speichervolumens sind diese kleineren Bilder besser für die Anzeige am Bildschirm geeignet, und es müssen bei diesem Format nicht zu viele Daten über das Netz transportiert werden. Deswegen eignet sich JPG für die Derivative.

Der Einfluss von Scanmodus, Auflösung und Speicherformat auf OCR

In diesem Schritt wird das Zusammenwirken der Einstellung bestimmter Werte bei Scanmodus, Auflösung und Speicherformat auf das OCR-Ergebnis betrachtet. Die folgende Tabelle fasst die Ergebnisse zusammen:

	schwarz-weiß		grau		farbig	
	TIFF	JPG	TIFF	JPG	TIFF	JPG
75 dpi	xx	xx	x	xx	xx	xx
100 dpi	x	xx	29 / 98,37 %	xx	15 / 99, 15 %	xx
200 dpi	8 / 99,55 %	26 / 98,54 %	6 / 99,66 %	x	7 / 99,61 %	x
300 dpi	9 / 99,49 %	9 / 99,49 %	8 / 99,55 %	16 / 99,10 %	10 / 99,44 %	19 / 98,93 %
600 dpi	7 / 99,61 %	9 / 99,49 %	9 / 99,49 %	14 / 99,21 %	11 / 99,38 %	11 / 99,38 %
Durchschnitt bei 300 und 600 dpi	99,55%	99,49 %	99,52 %	99,16 %	99,41 %	99,66 %

Tabelle 2: Ausgewählte Scaneinstellungen und ihr Einfluss auf das OCR-Ergebnis

Die Probescans wurden mit unterschiedlichen Auflösungswerten (linke Spalte: Werte von 75, 100, 200, 300 und 600 dpi), in drei Scan-Modi (schwarz-weiß, grau, farbig) und mit Verwendung von zwei Speicherformaten (TIFF – ohne Kompression, JPG – mit Kompression) durchgeführt. Der vor dem Schrägstrich aufgeführte Wert gibt die Anzahl der Fehler pro Seite an. Ein ‚x' bedeutet, dass die Seite schwer lesbar ist (Korrektur lohnt nicht), zwei ‚x' zeigen an, dass das OCR-Ergebnis nicht lesbar und nicht verwendbar ist. Nach dem Schrägstrich steht die Anzahl der richtig erkannten Zeichen als Prozentwert.

Die oben genannte Buchseite besteht aus 1 777 Zeichen (gezählt mit MS Word 2007), die mit dem OCR-Verfahren richtig erkannt werden sollten. Es werden nicht nur Buchstaben an sich bewertet, sondern auch Satz- und Leerzeichen zwischen den einzelnen Wörtern. Jedes nicht richtig erkannte Zeichen, das durch Korrektur neu eingetippt werden muss, wird als ein OCR-Fehler angesehen. Für die Berechnung der Durchschnittswerte werden nur die Einträge von 300 und 600 dpi berücksichtigt, weil nur in diesen Bereichen jedes Mal ein annehmbares OCR-Ergebnis entstand.

Um die Ergebnisse der optischen Zeichenerkennung zu verbessern, können zusätzliche Schritte unternommen werden, die nicht direkt mit dem Scannen verbunden sind. Einerseits können die gescannten Bilder mit einem Bildverarbeitungsprogramm so modifiziert werden, dass z. B. die Schärfe (dies war im Rahmen dieses Projektes nicht nötig) oder die Werte für Kontrast angepasst werden. Andererseits kann ein besseres

OCR-Ergebnis durch spezifische Einstellungen innerhalb der verwendeten OCR-Software (Spracheinstellungen und Benutzermuster) erreicht werden.

Bei den bearbeiteten Texten sind die Sprachen Deutsch und Tschechisch vertreten. Um die Buchstaben dieser beiden Sprachen korrekt zu erkennen, wenn z. B. im deutschen Text tschechische Zeichen vorkommen, wird zu dem Deutschen auch das Tschechische als zweite Dokumentsprache zusätzlich definiert.

Andere Sonderzeichen im Text bleiben nicht unbeachtet. Das Programm wird auf neue Muster trainiert, um Sonderzeichen oder schlecht erkennbare Zeichen besser zu identifizieren. Um diese Art der Zeichen eindeutig unterscheiden zu können, wird die Funktion ‚Benutzermuster testen' des verwendeten Programms ‚ABBY FineReader' eingesetzt. Ergänzend wird noch eine Probe mit der Verwendung des Lernmusters durchgeführt:

TIFF	schwarz-weiß	grau
100 dpi	x / x	29 / 28
300 dpi	9 / 2	8 / 1
600 dpi	7 / 1	9 / 2

Tabelle 3: OCR-Bearbeitung mit Verwendung eines Lernmusters

Die Tabelle Nr. 3 zeigt das Ergebnis der zweiten OCR-Bearbeitungsprobe. In diesem Schritt wurden nur noch die TIFF-Dateien, mit Auflösung 100, 300 und 600 dpi, schwarz-weiß und grau behandelt. Bei den Lese-Optionen von FineReader wurde im Mustereditor ein neues Muster eingelegt, und beim Lesen der Probedateien lernte das Programm, die neuen Zeichen zu unterscheiden.

Die Ergebniswerte bestehen jeweils aus zwei Angaben, die Ziffer links vom Schrägstrich stammt aus der früheren Probe (ohne Lernmuster), die Ziffer rechts vom Schrägstrich ist das Ergebnis der zweiten Probe (mit Lernmuster). Das „x" bedeutet, dass der Text sehr schwer lesbar und praktisch unverwendbar ist. Bei der Datei in Grau mit Auflösung 100 dpi, wurde nur eine geringe Verbesserung erzielt. Doch bei den Dateien mit Auflösung von 300 und 600 dpi ist eine wesentliche Verbesserung der OCR-Resultate zu erkennen.

Es ist empfehlenswert, bei der OCR-Bearbeitung die Funktion ‚Benutzermuster testen' zu verwenden. Im Anhang Nr. 2 werden die Buchstabenmuster und

Kodierungsvorgaben für die Texterfassung aufgeführt. Die verwendeten Texte weisen eine gute Qualität auf, es gab keine wesentlichen Schwierigkeiten in der Bearbeitung.

Nicht alle Bücher des Projekts mussten gescannt werden, wie es bei der tschechischen Version von Jiří Kratochvils ‚*Nesmrtelný příběh*' der Fall war. Dieses Werk hat der Verlag als eine Datei im PDF-Format zur Verfügung gestellt. Die Transformation dieses Formats in eine Textdatei erfolgte ebenfalls durch das OCR-Verfahren mit ABBY FineReader. Im nächsten Schritt werden noch die zwei ausgewählten Speicherformate getestet.

Speicherformat

Das verwendete Speicherformat nimmt unterschiedlich viel Speicherplatz in Anspruch. In dieser Arbeit werden die zwei bereits erwähnten Speicherformate JPG und TIFF kurz verglichen.

Die unten angegebene Tabelle bietet eine Übersicht der Speichergrößen der bei den Probescans bearbeiteten Dateien. Die Angaben (in Kilobytes / KB) wurden dem Dateimanager ‚Nautilus' entnommen:

	schwarz-weiß		grau		farbig	
	TIFF	JPG	TIFF	JPG	TIFF	JPG
75 dpi	19,3	11,5	149,1	4,7	446,9	5,5
100 dpi	33,8	18,3	264,4	8,0	792,8	9,7
200 dpi	133,3	50,7	1 055,0	28,4	3 166,0	33,4
300 dpi	298,7	91,3	2 375,0	54,0	7 128,0	64,8
600 dpi	1 190,0	257,9	9 504,0	54,9	28 503,0	204,2

Tabelle 4: Speichergröße der bearbeiteten Dateien in KB

In dieser Tabelle werden unterschiedliche Auflösungswerte (75, 100, 200, 300 und 600 dpi) in drei Scanmodi (schwarz-weiß, grau, farbig) und zwei Speicherformate (TIFF und JPG) in ihrem Zusammenspiel betrachtet. Diese Untersuchung vergleicht die Speichergröße der Dateien. Bei den kleineren JPG-Dateien wird eine verlustbehaftete Kompression verwendet. Im Gegensatz dazu bleiben bei den TIFF-Dateien alle Informationen erhalten, denn dieses Format verwendet eine verlustfreie Kompression (hier jedoch unkomprimiert). Die komprimierten Dateien brauchen zwar nicht so viel Speicherplatz, aber sobald sie beschädigt werden, ist es nicht mehr möglich, den Inhalt

wiederherzustellen. Deswegen ist es für die Dauerspeicherung die bessere Wahl, die Dateien unkomprimiert zu speichern.

Für die Masterdateien, die hier als Graustufenbilder realisiert werden, wird nach dem derzeitigen Kenntnisstand die Wahl ‚TIFF uncompressed' empfohlen. TIFF hat während seiner Existenz die wichtigsten Standards für Speicherung der Bildinformation etabliert, und es kann damit gerechnet werden, dass es auch in Zukunft von den gängigen Computerprogrammen unterstützt wird. Falls benötigt, kann aus Platzgründen eine verlustlose Kompression angewendet werden. Das Format JPG sollte nur für die abgeleiteten Kopien verwendet werden.[35] Der Vergleich dieser Werte kann bei der Planung des Speicherplatzes für das Projekt behilflich sein, obwohl für die geplante Verwendung der Dateien zuletzt die Bildqualität entscheidend ist. Aus den Tests können im nächsten Schritt die endgültigen Werte für die Bildbearbeitung in diesem Projekt gewählt werden.

Auswertung der Probescans

Die durchgeführte Analyse hilft die passenden Scanner-Einstellungen zu bestimmen, die für die im DeuCze-Projekt zu scannenden Bücher verwendet werden:

a) Scanmodi – die Wahl zwischen den Scanmodi (Farbenmodi) beschränkt sich hiermit auf die Entscheidung zwischen „grau" und „schwarz-weiß". Die Einstellung „farbig" bringt anscheinend keine wesentliche Verbesserung bei der OCR-Verarbeitung und die Dateien nehmen unnötig viel Speicherplatz ein. Bei diesem Projekt handelt es sich um moderne Bücher, deren Druck keine Alterungsspuren aufweist. Gleichzeitig kommen in den verwendeten Texten keine Elemente vor, die durch unterschiedliche Farben bestimmte Information übertragen würden, wie es beispielsweise in der Kartografie der Fall ist. Aus den genannten Gründen wird im Weiteren auf die Verwendung von Farbbildern verzichtet.

Bei Bedarf können von den Quelldaten die schwarz-weißen Bilder erstellt werden (eine Konversion von schwarz-weiß zu grau ist nicht mehr möglich), deswegen sind die grauen Bilder besser für die Archivierung geeignet. Daneben wird auch die Tatsache positiv bewertet, dass die in grau verfassten Dateien mit kleinerer Auflösung für die Benutzer besser lesbar sind als die schwarz-weißen Bilder. Die Kopien für die

35 Vgl. DFG-Praxisregeln „Digitalisierung", S. 8f.

Benutzeroberfläche, die eine kleinere Auflösung haben und komprimiert sind (Speicherformat JPG), vermitteln einen schnelleren Zugang (es wird eine kleinere Menge an Daten über das Netz übertragen) in einer ausreichenden Qualität (vgl. KENNEY 2000, 55f).

b) die Auflösung – bei der Auflösungsbestimmung werden die OCR-Ergebnisse mit Berücksichtigung der vom Scanner angebotenen Auflösungswerte betrachtet. Die Qualität der OCR-Ergebnisse, wo die Werte 75, 100 und 200 dpi durchschnittlich keine zufriedenstellenden Ergebnisse lieferten, schränkt die Auswahl auf die Werte von 300 und 600 dpi ein. In den verwendeten Texten kommen Wörter aus anderen Sprachen vor, und mithilfe der höheren Auflösung kann das OCR-Programm die diversen Sonderzeichen besser entschlüsseln. Gleichzeitig ist es günstiger, die Masterdateien eher in besserer Qualität zu speichern, von denen bei Bedarf jederzeit Kopien in gewünschter Qualität gebildet werden.

c) das Speicherformat – die Wahl zwischen den Speicher-Formaten TIFF und JPG beeinflusst Tatsachen wie: den benötigten Speicherplatz, die verwendete Kompression oder die Möglichkeit, Metadaten mitzuspeichern. Im Rahmen dieses Projektes wird mit zwei Bilddateiarten gearbeitet. Auf der einen Seite sind das die Masterdateien, die durch das Scannen gewonnen wurden und in der höheren Auflösung von 600 dpi abzuspeichern sind; für diese ist das TIFF-Format geeignet, und sie werden zu Archivzwecken und in Kopien für die OCR-Bearbeitung verwendet. Auf der anderen Seite handelt es sich um kleinere JPG-Dateien, die für die Benutzeroberfläche bestimmt sind.[36]

Die Dateinamen beim Speichern der Masterdateien bestehen aus Informationen zum jeweiligen Werk, z. B. bei dem Kratochvil-Text (`KRN05_006_ cz.tif`): das Kürzel `KRN` gibt durch die ersten zwei Buchstaben den Autorennamen (`KR` = Kratochvil) und den ersten Buchstaben des Werktitels (`N` = ‚Nesmrtelný příběh') an. Es folgen unmittelbar zwei Ziffern, die das Erscheinungsjahr des Werkes angeben (`05` = das Erscheinungsjahr 2005). Nach einem Unterstrich folgt die Nummer der Datei (stimmt mit Buchseite überein), und die letzte Angabe nach dem zweiten Unterstrich ist die

36 Die Masterdateien nehmen im Durchschnitt je Bild 12 MB Speicherplatz ein, bei den Derivativen sind es im Durchschnitt jeweils 170 KB.

Sprache des Werkes (cz = Tschechisch); die Dateierweiterung .tif nach dem Punkt gibt an, dass die Datei im TIFF-Format gespeichert ist.

Die Namen der Derivative sind ähnlich konzipiert, z. B. GRU92_105_de.jpg bezeichnet den Autor Günter Grass (GR), das Werk ‚Unkenrufe' (U) und das Erscheinungsjahr 1992 (92). Die darauf folgende Nummer gibt die Seitenzahl der dargestellten Seite (hier die Seite 105), und die Sprache ist Deutsch (de). Die Dateierweiterung .jpg gibt an, dass es sich um eine abgeleitete Datei im JPG-Format handelt.

2.5 Zusammenfassung

Dieses Kapitel ist den Schritten auf dem Weg vom gedruckten Text bis zu einer Textdatei gewidmet. Einleitend wurden die zu bearbeitenden Bücher vorgestellt. Im zweiten Unterkapitel wurden die allgemeinen Kriterien für die Qualität der Scans behandelt. Es werden zwei Typen der Bilddateien unterschieden, einerseits die qualitativ hochwertigen Bilder, sog. Masterdateien (‚digital master'), die vor allem den Archivzwecken dienen. Von diesen Dateien werden Derivative abgeleitet, die der weiteren Bearbeitung und der Verwendung für die Anzeige im Korpus dienen.

Im dritten Unterkapitel wurden die Scanner-Einstellungen (wie Farbeinstellungen und Auflösung) vorgestellt. Einen untrennbaren Teil der Digitalisierung der Bilder stellt die Qualitätskontrolle dar, um die eventuell entdeckten Mängel auszubessern.

Das abschließende Unterkapitel behandelte die konkreten Scanner-Einstellungen. Damit das Gewinnen der Texte mit möglichst kleinem Aufwand realisiert werden kann, wurden Probescans durchgeführt und so die optimalen Scaneinstellungen bestimmt. Die digitalisierten Buchseiten wurden schließlich mit der optischen Zeichenerkennungssoftware bearbeitet. In dieser Phase stehen die Texte für die weitere Bearbeitung innerhalb des Projektes bereit.

3 Das Erstellen der XML-Datei

Die unformatierte Textdatei, die durch das OCR-Verfahren entstanden ist, wird in eine wohlgeformte und gültige XML-Datei transformiert.[37] Die zum Schluss erstellte Datei dient als Grundlage für das bearbeitete Korpus.

3.1 XML

Für die Speicherung der Korpus-Dateien wird das XML-Format verwendet. Die Auszeichnungssprache XML (englisch ‚eXtensible Markup Language') ist als ein Standard für die Speicherung und Bearbeitung der Textdokumente geeignet (vgl. Kosek 2000, 9ff). Die XML-Markierung beschreibt hier überwiegend die logische Struktur der Dokumente und ist für den Endbenutzer nicht sichtbar. Es handelt sich um ein offenes textbasiertes Format, das von vielen Text-Verarbeitungsprogrammen bearbeitet werden kann, und es ist somit an kein bestimmtes Betriebssystem oder an keine bestimmte Software gebunden. Dieses Format kann mit gängigen Browsern angezeigt werden – die Darstellung der Texte im gewünschten Layout vermittelt die Stylesheet-Sprache CSS (englisch ‚Cascading Style Sheet').

Das Speichern in proprietären Formaten wurde abgelehnt, weil es vorkommen kann, dass die jeweilige Software in Zukunft nicht mehr unterstützt wird, und so wären die Daten so gut wie verloren.

Das XML-Dokument hat eine hierarchische Struktur und besteht aus ineinander verschachtelten Einheiten (sog. Elementen), die durch die Markierungen, die Tags[38], gekennzeichnet werden. Die meisten Tags kommen in Paaren vor, d. h., sie bezeichnen jeweils den Anfang (Starttag) und das Ende (Endtag) des gegebenen Elements, z. B. Absatzmarkierung: … (vgl. McEnery / Xiao / Tono 2006, 22f). Für die Kennzeichnung bestimmter Phänomene (hier lineare Strukturen wie Zeilen- und Seitenumbruch) werden Elemente verwendet, die keinen Inhalt haben (sog. ‚empty tags') und die direkt im Starttag geschlossen werden, z. B. Zeilenumbruch:
.

Den Elementen können bestimmte Attribute beigefügt werden, die die Bedeutung der Elemente präzisieren, wie z. B. Nummerierung der Absätze: <absatz

37 Im Grunde genommen ist die XML-Datei auch eine Textdatei. Die Bezeichnung ‚Textdatei' wir hier im Sinne einer Textdatei verwendet, die keine Formatierung enthält (sog. ‚plain text file'). Die ‚XML-Datei' wird hier als Textdatei mit einer vollständigen XML-Markierung interpretiert.

38 ‚das Tag' – ein Tag im Sinne der Markierung ist ein Neutrum.

nummer="1">...</absatz> (vgl. KOSEK 2000, 26f). Das Gesamtdokument steht in einem einzigen Wurzel-Element, und alle weiteren Elemente sind entsprechend verschachtelt, wie z. B. die Kodierung für einen Satz innerhalb eines Absatzes: <s>…</s> . Eine fehlerhafte Markierung liegt vor, wenn sich die Elemente überlappen, z. B. wenn das Satztag erst nach dem Absatztag geschlossen wird (<s>… </s>). Ein korrekt zusammengestelltes Dokument wird als ‚wohlgeformt' bzw. ‚well-formed' bezeichnet (vgl. KOSEK 2000, 26f).

Die Elementnamen sind im Rahmen von XML nicht festgelegt, sie können beliebig gewählt werden. Aus Gründen der Übersichtlichkeit und um den Datenaustausch zu vereinfachen, werden oft vordefinierte Tagsets benutzt. Dieses Projekt lehnt sich an die Markierungsregeln von ‚The Text Encoding Initiative' (TEI)[39] an.

Um zu testen, ob die XML-Datei den Regeln entspricht, wird sie gegen ein Schema validiert. Im verwendeten Schema wird angegeben, welche Elemente und Attribute im jeweiligen Dokument vorkommen können und wie sie miteinander kombiniert werden dürfen (vgl. KOSEK 2000, 37ff).

Es hängt von dem geplanten Verwendungszweck ab, bis zur welchen Stufe der Textstruktur die Markierung im jeweiligen Projekt verzeichnet wird. Das vorgesehene Ziel in dieser Phase ist die Synoptisierung der Absätze und Sätze in den deutschen und tschechischen Texten[40]. In der Zieldatei wird die Textstruktur bis zur Satzebene getaggt, sodass der Text bei der Anzeige in der Benutzeroberfläche des Korpus nach Sätzen synoptisiert angezeigt werden kann. Die eigentliche Umwandlung der Textdatei zu der XML-Datei erfolgt mithilfe der Programmiersprachen PHP und XSLT.

3.2 Das System der Tags

In diesem Unterkapitel werden die einzelnen Tags und ihre Verwendung in diesem Projekt beschrieben[41]. Die TEI stellt weit mehr Tags zur Kodierung der Korpustexte zur Verfügung. In diesem Abschnitt werden jedoch nur die Tags behandelt, die bei der Kodierung der DeuCze-Texte bisher tatsächlich verwendet wurden. Wenn in das

39 Vgl.: http://www.tei-c.org, zit.: 13. 7. 2009.

40 Das Synoptisierungsskript wird in dieser Arbeit nicht näher behandelt, da es von anderen Projektmitgliedern gebaut wurde.

41 Die Tagbeschreibung geht strukturell von dem Dokument „Leitlinien zur Kodierung des Campe-Wörterbuchs in TEI P5" von Mirjam Blümm aus. Zugänglich unter WWW: http://www.textgrid.de/fileadmin/TextGrid/veroeffentlichungen/Leitlinien-zur-Kodierung-des-Campe-Woerterbuchs-in-TEI-P5.pdf, zit.: 20. 2. 2010.

Korpus neue Texte mit neuen Elementen (z. B. eine E-Mail) aufgenommen werden, müssen entsprechende Tags aus dem TEI-Satz ergänzt werden, und so wird dieses System der Tags für DeuCze erweitert. Weitere Tags müssen auch dann geholt werden, wenn z. B. die Beschreibung der Wortarten in die bestehenden Texte eingefügt werden soll. Das System der Tags für das DeuCze-Korpus wird also zwar als eine Gesamtheit der Tags beschrieben, dies bedeutet jedoch nicht, dass es abgeschlossen und nicht mehr weiter aufnahmefähig ist.

Der Text soll schließlich nach Sätzen segmentiert werden, damit die satzweise parallele Anzeige der beiden vertretenen Sprachen möglich ist. Der Satz wird jedoch nicht im grammatischen Sinne bestimmt, sondern nach typografischen Regeln, wie z. B. dass ein Satz mit Großbuchstaben anfängt und mit einem Satzschlusszeichen endet.

Wenn es in weiteren Auslegungen nicht anders angegeben wird, geht die nachfolgende Tag-Übersicht von der deutschen Fassung des Buches von Jiří Kratochvil aus. Sollte es Unterschiede bei der tschechischen Fassung des Kratochvil-Textes oder der Texte von Thomas Brussig geben, wird dies konkret angegeben. In den Anhängen befinden sich ausgewählte Vertreter der hier beschriebenen Dateien und Skripte. Aus räumlichen Gründen können nicht alle Dateien des Projekts angeführt werden. Die vertretenen Dateien sind gekürzt, um nur das Wesentliche zu illustrieren. Die vollständigen Dateien sind im Archiv des Projektes auf dem Server der Würzburger Arbeitsstelle gespeichert.

Die hier verwendete Markierung und Strukturierung der Korpustexte entspricht der XML-Struktur nach den Regeln von TEI P5, d. h., dass bestimmte Textabschnitte mit entsprechenden XML-Markierungen (XML-Tags) versehen werden[42].

Bei der Kodierung der Informationen in Korpustexten wird nicht nur die XML-Struktur verwendet, als Beispiel für eine andere Vorgehensweise sei das Brown-Corpus genannt. In diesem Korpus werden die Wortarten mithilfe unterschiedlicher Buchstabenkombinationen gekennzeichnet, z. B. NN für Substantiv.[43] Für den Bedarf dieses Projektes sollte auch die Strukturierung der Texte (wie Absatz oder Seite im Quelltext) im Korpustext enthalten sein, damit die Belege entsprechend zitiert werden können.

42 Die Elemente bestehen jeweils aus einem Anfang- und Endtag, die Elemente sind hierarchisch angeordnet und durch Attribute können weitere Beschreibungsmerkmale ergänzt werden; mehr dazu vgl.: „P5: Guidelines for Electronic Text Encoding an Interchange“ im Kapitel „Gentle Introduction to XML“ (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html, zit.: 13. 7. 2009.).

43 Vgl. http://icame.uib.no/brown/bcm.html, zit.: 23. 11. 2010.

Für dieses Projekt wurde das XML-Format auch deswegen gewählt, weil für das DeuCze-Korpus ein großer Wert auf die Standardisierung des Aufbereitungsprozesses und der Verwendung der Texte gelegt wird. Eine XML-Datei ist keine binäre Datei an sich, sondern eigentlich eine Textdatei und ist auch für den Menschen lesbar und verständlich. Sie ist von keinem konkreten Programm abhängig und kann auf verschiedenen Betriebssystemen gelesen und bearbeitet werden. Diese Tatsachen erleichtern nicht nur den Datenaustausch zwischen den Projektteilnehmern, sondern es wird angenommen, dass diese Dateien unter diesen Bedingungen auch in Zukunft lesbar sind.[44]

Das XML-Dokument muss wohlgeformt und valide sein. Wohlgeformt ist es, wenn die Elemente korrekt verschachtelt sind und die hierarchische Struktur nicht gestört wird. Das TEI-Dokument kann erst dann als valide bezeichnet werden, wenn es mindestens aus den Teilen ‚header' und ‚text' besteht (vgl. McEnery / Wilson ^{2}2001, 35). Der Header enthält die bibliografischen Angaben zu dem Text und die Beschreibung der Kodierung. Es können auch Angaben zum Textprofil (z. B. Dokumentsprache) und eine Korrekturübersicht (Korrektur der Satzfehler) enthalten sein. In diesem Abschnitt werden nur die Tags behandelt, die zur unmittelbaren Kodierung der Quelltexte verwendet werden. Dieses Tagset fasst die hier verwendeten Tags zusammen. Werden in das Korpus neue Bücher mit neuen Phänomenen (wie z. B. eine E-Mail) aufgenommen, werden entsprechende Tags ergänzt, jeweils aber in Übereinstimmung mit den TEI-Regeln.

Die einzelnen Tags werden mithilfe eines PHP-Skripts in den Roman-Text eingefügt und mit einer XSLT-Schablone nummeriert und zusätzlich angepasst. Damit keine Elemente aus dem Header zusammen mit den Text-Elementen nummeriert werden, wird der Header erst im letzten Schritt der Erstellung der XML-Datei manuell eingefügt. Die vollständige XML-Datei wird schließlich gegen ein Schema validiert, das die Regeln für die jeweilige XML-Struktur angibt. Das hier verwendete Schema ist in der Sprache RelaxNG verfasst und wird mit dem Instrument ROMA[45], das über die TEI-Webseite zugänglich ist, automatisch erstellt. Die Tags werden je nach dem Verwendungsgebiet in Modulen gruppiert, für die Bedürfnisse dieses Projektes enthält

44 Im Vergleich zu proprietären Formaten und spezifischen Beschreibungsweisen.
45 Siehe: http://www.tei-c.org/Roma, zit.: 13. 7. 2009.

das Schema diese Module: ‚core', ‚tei', ‚header', ‚textstructure', ‚linking' und ‚analysis'. Aus den Modulen werden jene Elemente, die keine Verwendung in diesem Projekt finden, ausgelassen. Die Einstellungen für das mit ROMA erstellte Schema werden in einer Datei, die die Benutzereinstellungen enthält (die sog. Customisation-Datei) gespeichert. Diese Datei kann bei Bedarf wieder in das ROMA-Tool hochgeladen werden, um das Schema so den aktuellen Bedürfnissen des Projektes anzupassen.

Manche der verwendeten Elemente werden durch die Verwendung ausgewählter Attribute näher charakterisiert. In den Elementen, die für die Kodierung des Textes innerhalb dieses Projektes verwendet wurden, kommen folgende Attribute vor:

- `n` – trägt den Wert der Seiten- und Zeilennummerierung
- `rend` – beschreibt das jeweilige Element näher, wie es im Quelltext abgebildet ist, z. B. bestimmt es die Hervorhebungsart, wie Kursiv-, Fett- oder Kapitälchen-schrift
- `type` – dient hier der Unterscheidung zwischen den Buch- und Kapitelüberschriften
- `xml:id` – eindeutige und einmalige Identifizierung der jeweiligen Textpassage und Referenz im Quelltext. Z. B. gibt der Attributwert `div1_jku00de` an, dass es sich um das erste `<div>`-Element handelt. Nach dem Unterstrich folgt die Autorenidentifizierung (hier `jku00de`). Diese Abkürzung gibt den Autorennamen (`jk` steht für Jiří Kratochvil), das Buch (`u` steht für ‚Unsterbliche Geschichte'), das Erscheinungsjahr (`00` steht für das Jahr 2000) und die Sprache (`de` bedeutet ‚Deutsch') an.

Alle Attribute mit ihrem konkreten Wert werden später bei der Beschreibung der jeweiligen Elemente näher beschrieben.

Die im Text getaggten Phänomene beschreiben:

a) die hierarchische Struktur des Romans:

- Bücher – als Teile des Romans (nur bei Kratochvil) + Überschriften
- Kapitel + Überschriften
- Absätze
- Segmente und Sätze
- Sonderstrukturen wie der Brief (der Brief nur bei Kratochvil) oder das Gedicht

b) die typografischen Eigenschaften des Textes:

- Hervorhebung im Text (Kursiv-, Fett- und Kapitälchenschrift) – kann auf unterschiedlichen Stufen der Hierarchie vorkommen

c) die linearen Eigenschaften des Textes:

- Seiten- und Zeilenumbruch

Die XML-Struktur ist streng hierarchisch geordnet. Das gesamte Dokument befindet sich im Wurzelelement `<TEI> … </TEI>` und teilt sich in zwei grundlegende Teile: ‚TEI-Header' und ‚Text'. Im Text-Teil steht das Element `<body>`, das den gesamten Romantext enthält. Das unten wiedergegebene Bild zeigt die Elementstruktur des Kratochvil-Textes an:

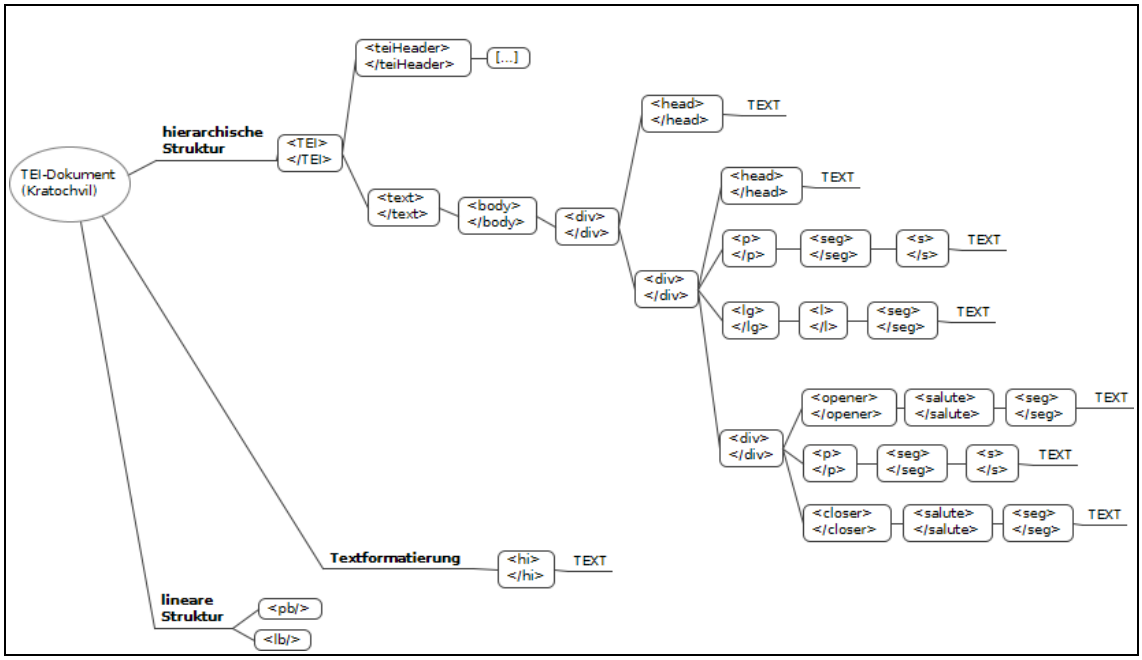


Bild 7: Schematische Übersicht der verwendeten TEI P5-Elemente

Auf dem Bild Nr. 7 sind die TEI P5-Elemente schematisch abgebildet, so wie sie im Kratochvil-Text verwendet werden. Die Posten mit ‚TEXT' weisen darauf hin, dass keine Elemente an sich mehr enthalten sind, sondern bereits der Text der Vorlage. Auf dem Bild werden keine Attribute angeführt, diese werden bei der Beschreibung der konkreten Elemente angegeben und erklärt. Gleichzeitig wird der Übersichtlichkeit wegen auch der Header (das Element ‚teiHeader') gekürzt abgebildet. Die Elemente, die unter ‚Textformatierung' und ‚lineare Struktur' stehen, können auf unterschiedlichen Ebenen der hierarchischen Struktur vorkommen.

Jede valide TEI-Datei enthält eine Beschreibung ihres Inhalts, die am Anfang der Datei im Header (`<teiHeader>...</teiHeader>`) steht. Die Aufteilung des Headers ist:

1) die Dateibeschreibung (‚file description') steht in den Tags `<fileDesc>` … `</fileDesc>` und enthält die bibliografischen Angaben zu dem kodierten Text. Dieses Element ist im Header obligatorisch.

2) die Beschreibung der Kodierung (‚encoding description') steht in den Tags `</encodingDesc>...</encodingDesc>` und enthält Informationen dazu, wie und zu welchem Zweck der jeweilige Text kodiert wird.

3) das Textprofil (‚text profile') wird durch die Elemente `<profileDesc>` … `</profileDesc>` geklammert. Es sind Kontextinformationen zu dem Text enthalten, wie z. B. die Sprache des Textes.

4) die Korrekturenübersicht (‚revision history') listet innerhalb der Tags `<revisionDesc>` … `</revisionDesc>` die im Originaltext durchgeführten Satzfehlerkorrekturen auf.[46]

In den folgenden Unterkapiteln werden die einzelnen Elemente und Attribute, im Bezug auf ihre Verwendung in diesem Projekt, beschrieben. In der Regel wird bei der Beschreibung so vorgegangen, dass zuerst das Element an sich vorgestellt wird, dann werden Beispiele und jeweils abschließend die Attribute behandelt.

3.2.1 `<div>`

→ ‚text division' / Textdivision – der durch `<div>` … `</div>` geklammerte Textabschnitt bedeutet eine Subdivision innerhalb von `<body>`. Der Roman von Kratochvil teilt sich in fünf Teile (Bücher). Die einzelnen Roman-Bücher, die Kapitel und der (nur bei Kratochvil) vorkommende Brief werden jeweils mit `<div>`-Elementen markiert. Die `<div>`-Arten werden folgendermaßen unterschieden[47]:

a) - durch das Ergänzen des Attributs `type` mit dem Wert `book` wird das Buch als Teil des Romans markiert. Folgendes Beispiel zeigt die Markierung des ersten Buches im Roman an:

```
<div type="book">
<pb n="7"/><lb n="7:1"/>
  <head>Erstes Buch<lb n="7:2"/>DER CHIMPANSE
```

46 Die komplette Header-Beschreibung befindet sich im PDF-Dokument ‚TEI P5: Guidelines for Electronic Text Encoding and Interchange', S. 17ff. Zugänglich unter WWW: http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf

47 Wegen dem Umfang der Beispiele, werden bei dem `<div>`-Element keine Bilder, nur die gekürzte Kodierung, angeführt.

```
  </head>
<pb n="8"/>
  <div type="chapter" xml:id="div1_jku00de">
<pb n="9"/><lb n="9:1"/>
    <head>1
<lb n="9:2"/>Die Stimme meines Herrn
    </head>
<lb n="9:3"/>
    <p xml:id="div1.p1_jku00de">
      <seg xml:id="div1.p1.seg1_jku00de">
        <s xml:id="div1.p1.seg1.s1_jku00de">
          Geboren wurde ich, wenn Sie das wirklich hören wol-
<lb n="9:4"/>len, meine Herren, geboren wurde ich in der Nacht
vom
<lb n="9:5"/>31. Dezember 1899 auf den 1. Januar 1900, und mein
Vater

  […]

<lb n="79:23"/>zerstörten Lokomotive, bösen Träumen und von der
Bewunde-
<lb n="79:24"/>rung für alle gütigen Monarchen für immer geheilt
von der ga-
<lb n="79:25"/>lizischen Front zurück.
        </s>
      </seg>
    </p>
  </div>
</div>
```

b) - durch das Ergänzen des Attributs `type` mit dem Wert `chapter` wird das `<div>`-Element als Kapitel des Romans spezifiziert:

```
<div type="chapter" xml:id="div1_jku00de">
<pb n="9"/><lb n="9:1"/>
  <head>1
<lb n="9:2"/>Die Stimme meines Herrn
  </head>
<lb n="9:3"/>
  <p xml:id="div1.p1_jku00de">
    <seg xml:id="div1.p1.seg1_jku00de">
      <s xml:id="div1.p1.seg1.s1_jku00de">
        Geboren wurde ich, wenn Sie das wirklich hören wol-
<lb n="9:4"/>len, meine Herren, geboren wurde ich in der Nacht
vom
<lb n="9:5"/>31. Dezember 1899 auf den 1. Januar 1900, und mein
Vater

  […]

<lb n="79:23"/>zerstörten Lokomotive, bösen Träumen und von der
Bewunde-
<lb n="79:24"/>rung für alle gütigen Monarchen für immer geheilt
von der ga-
<lb n="79:25"/>lizischen Front zurück.
      </s>
```

```
    </seg>
  </p>
</div>
```

Das zweite verwendete Attribut (`xml:id`) dient der eindeutigen Identifizierung des Elements, bei dem es verwendet ist, und gibt die Textreferenz an.

c) - in einem Fall kommt `<div>` ohne weitere Attribute vor. Auf diese Weise wird eine Subdivision im Text (ein Brief, der nur im Kratochvil-Text vorkommt) markiert:

```
<div>
  <opener>
    <salute xml:id="div33.sal424_jku00de">
      <seg xml:id="div33.p424.seg1327_jku00de">
      <hi rend="italic">Teuerster Freund!</hi>
      </seg>
    </salute>
  </opener>
  <lb n="161:23"/>
  <p xml:id="div33.p425_jku00de">
    <hi rend="italic"><seg xml:id="div33.p425.seg1328_jku00de">
      <s xml:id="div33.p425.seg1328.s1400_jku00de">
        Ich schreibe nur noch schlecht und mit Mühe, aber gerade
deshalb ist es

  […]

      <lb n="162:27"/>irre), und nehmen Sie bitte meinen Dank
entgegen.
      </s>
    </seg></hi>
  </p>
  <lb n="162:28"/>
  <closer>
    <salute xml:id="div33.sal426_jku00de">
      <seg xml:id="div33.p426.seg1351_jku00de">
        <hi rend="italic">Ihr
        <lb n="162:29"/>Thomas G. M.</hi>
      </seg>
    </salute>
  </closer>
</div>
```

Im Text von Brussig findet keine Teilung des Romans in Bücher statt, sondern er wird direkt in Kapitel geteilt.

3.2.2 `<head>`

→ ‚heading' / die Überschrift. Das paarige Tag `<head> … </head>` markiert die Überschriften. Im Kratochvil-Text kommen jeweils Buch- und Kapitelüberschriften vor.

Die Unterscheidung zwischen Buch und Kapitel wird nur innerhalb des `<div>`-Elements durch das `type`-Attribut angegeben; bei dem Element `<head>` werden die Überschriften nicht mehr differenziert und es werden keine Attribute ergänzt.

Der Titel des Buches steht auf einer Seite selbstständig ohne weiteren Text:


```
<head>Erstes Buch
<lb n="7:2"/>DER CHIMPANSE</head>
```

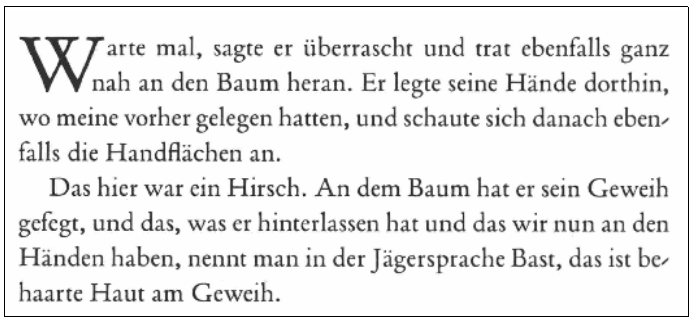
Die Kapitelüberschrift steht jeweils direkt vor dem nachfolgenden Kapiteltext:


```
<head>1
<lb n="9:2"/>Die Stimme meines Herrn
</head>
```

3.2.3 <p>

→ ‚paragraph' / Absatz – die Tags `<p>` … `</p>` markieren die Absätze im Text.

Jeder erste Absatz im jeweiligen Kapitel beginnt mit einer Initiale (im Beispiel ist es der erste Absatz) und jeder weitere Absatz wird links eingezogen (im Beispiel ist es der zweite Absatz). Diese Information wird nicht mehr bei jedem einzelnen Absatz festgehalten, sondern nur innerhalb des Headers im Teil zur Beschreibung der Kodierung (‚encoding description') verzeichnet.


Warte mal, sagte er überrascht und trat ebenfalls ganz nah an den Baum heran. Er legte seine Hände dorthin, wo meine vorher gelegen hatten, und schaute sich danach ebenfalls die Handflächen an.

Das hier war ein Hirsch. An dem Baum hat er sein Geweih gefegt, und das, was er hinterlassen hat und das wir nun an den Händen haben, nennt man in der Jägersprache Bast, das ist behaarte Haut am Geweih.

```
<p xml:id="div20.p171_jku00de">
  <seg xml:id="div20.p171.seg564_jku00de">
    <s xml:id="div20.p171.seg564.s595_jku00de">
      Warte mal, sagte er überrascht und trat ebenfalls ganz
      <lb n="86:4"/>nah an den Baum heran.
    </s>
```

```
    <s xml:id="div20.p171.seg564.s596_jku00de">
      Er legte seine Hände dorthin,
      <lb n="86:5"/>
      wo meine vorher gelegen hatten, und schaute sich danach
eben-
      <lb n="86:6"/>
      falls die Handflächen an.
    </s>
  </seg>
</p>
<lb n="86:7"/>
<p xml:id="div20.p172_jku00de">
  <seg xml:id="div20.p172.seg565_jku00de">
    <s xml:id="div20.p172.seg565.s597_jku00de">
      Das hier war ein Hirsch.
    </s>
  </seg>
  <seg xml:id="div20.p172.seg566_jku00de">
    <s xml:id="div20.p172.seg566.s598_jku00de">
      An dem Baum hat er sein Geweih
      <lb n="86:8"/>
      gefegt, und das, was er hinterlassen hat und das wir nun an
den
      <lb n="86:9"/>
      Händen haben, nennt man in der Jägersprache Bast, das ist
be-
      <lb n="86:10"/>haarte Haut am Geweih.
    </s>
  </seg>
</p>
```

Der Wert des Attributs `xml:id` wird um die Angabe erweitert, dass es sich hier z. B. um den 171. Absatz im Gesamttext handelt (`p171`).

3.2.4 <s>

→ ‚s-unit' / satzartiger Textabschnitt – `<s> … </s>`. Im Vergleich zu den manuell markierten Absatzanfängen muss die Bestimmung der Satzgrenze wegen der großen Anzahl der Sätze im Korpus möglichst automatisch verlaufen. Der Beschreibung dieses Elements und der Bestimmung der Satzgrenze wird mehr Aufmerksamkeit gewidmet als den anderen Elementen, weil spezifische Phänomene behandelt werden müssen, die nur im jeweiligen Text vorkommen, wie z. B. Punkte, die nicht als Satzendepunkte, sondern als Abkürzungspunkte verwendet werden. Auf eine gründliche Satzmarkierung wird großer Wert gelegt, denn die Segmentierung der Texte auf der Satzebene dient der parallelen Anzeige der Texte im Korpus.

Beim Inhalt des `<s>`-Elements handelt es sich nicht unbedingt um einen grammatischen Satz. Das Verständnis eines Satzes für den Bedarf der automatischen

Bearbeitung geht von den typografischen Eigenschaften aus, d. h., ein Satz beginnt mit Großbuchstaben und endet mit Satzendezeichen. Ein Satzendezeichen kann ein Punkt bzw. Frage- oder Ausrufezeichen oder drei Punkte (als Auslassungspunkte) sein.

Nicht alle Punkte stellen Satzendepunkte dar, sie kommen auch in anderen Funktionen vor:

a-1) Punkt nach Ordinalzahl: *Seit dem 21. August*; *4. Januar 1900*

a-2) Punkt nach Ordinalzahl, die als römische Zahl geschrieben wird: *Zar Nikolaus ii.*, *iii. Klasse*

b) Abkürzungspunkt: *k.u.k.-Amtes*; *c. k.*; *o.k.*; *T.G. Masaryk*

c) Auslassungspunkte: *Mögen Sie recht behalten, Sergeant ... Trotzkij.*

Die Auslassungspunkte werden als drei Punkte (...) und nicht als ein Zeichen (…) behandelt, erst zum Schluss der Bearbeitung mit dem PHP-Skript werden die Auslassungspunkte in ein einzelnes Zeichen (…) umgewandelt.

d) Sonderfälle: es handelt sich zwar um ein Satzende, aber nach dem Punkt folgt ein Zusatzzeichen, wie z. B. eine schließende Klammer oder eine auf der Benutzeroberfläche nicht sichtbare Formatierungs-Markierung. Das Satzende-Tag (`</s>`) wird erst nach der Klammer eingesetzt:

Jetzt, wo wir im Bilde sind, kommt es auf das Wort auch nicht an.)[48]

bzw. nach dem Formatierungstag:

Versuchsstation des Weltuntergangs.</hi>[49]

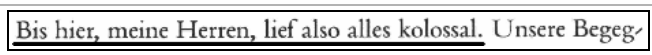
Nach der Behandlung aller Ausnahmen im Text kann davon ausgegangen werden, dass die übrig gebliebenen Punkte nur noch die Satzgrenzen bezeichnen. Eine zusammenfassende Übersicht der Bestimmungsarten der Satzgrenze bei den möglichen Satzendezeichen findet sich im Anhang Nr. 3.

Die Vorgehensweise bei der Markierung der Sätze wird im Unterkapitel „3.3.2 Die eigentliche Transformation in das XML-Format“, unter Punkt a) näher beschrieben, wo die Bearbeitung mit Hinweisen auf konkrete Stellen im Skript dargestellt wird. Die

48 Die öffnende Klammer steht weiter vorne im Quelltext.
49 Das Anfangstag steht weiter vorne im Quelltext.

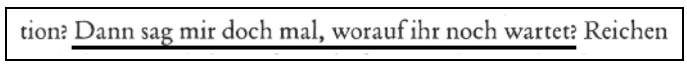
unten angeführten Beispiele[50] zeigen die Kodierung-Sätze, die nach Art der Satzendezeichen voneinander unterschieden werden:

a) ein Satz beginnt mit einem Großbuchstaben und endet mit einem Punkt, nach dem ein Großbuchstabe folgt:


Bis hier, meine Herren, lief also alles kolossal. Unsere Begeg-

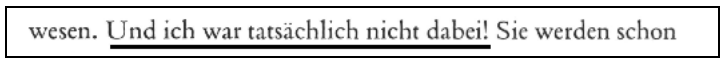
```
<seg xml:id="div1.p5.seg17_jku00de">
<s xml:id="div1.p5.seg17.s17_jku00de">
  Bis hier, meine Herren, lief also alles kolossal.
</s>
</seg>
<seg xml:id="div1.p5.seg18_jku00de">
<s xml:id="div1.p5.seg18.s18_jku00de">Unsere Begeg-
```

b) ein Satz beginnt mit einem Großbuchstaben und endet mit einem Fragezeichen, nach dem ein Großbuchstabe folgt:


tion? Dann sag mir doch mal, worauf ihr noch wartet? Reichen

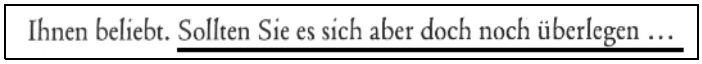
```
<seg xml:id="div36.p478.seg1516_jku00de">
<s xml:id="div36.p478.seg1516.s1597_jku00de">
  Dann sag mir doch mal, worauf ihr noch wartet?
</s>
</seg>
<seg xml:id="div36.p478.seg1517_jku00de">
<s xml:id="div36.p478.seg1517.s1598_jku00de"> Reichen
```

c) ein Satz beginnt mit einem Großbuchstaben und endet mit einem Ausrufezeichen, nach dem ein Großbuchstabe folgt:


wesen. Und ich war tatsächlich nicht dabei! Sie werden schon

```
<seg xml:id="div38.p490.seg1559_jku00de">
<s xml:id="div38.p490.seg1559.s1640_jku00de">
   Und ich war tatsächlich nicht dabei!
</s>
</seg>
<seg xml:id="div38.p490.seg1560_jku00de">
<s xml:id="div38.p490.seg1560.s1641_jku00de">Sie werden schon
```

d) ein Satz beginnt mit einem Großbuchstaben und endet mit drei Punkten / Auslassungspunkten (als Satzgrenze nur am Ende eines Absatzes):

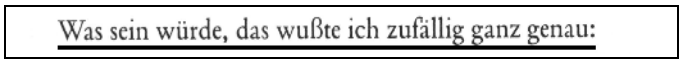

Ihnen beliebt. Sollten Sie es sich aber doch noch überlegen …

```
<s xml:id="div32.p408.seg1279.s1350_jkn05cz">
```

50 In manchen Fällen wird auch die textuelle Umgebung abgebildet (z. B. eine vollständige Zeile), damit die Satzgrenze deutlich zu beobachten ist. Das behandelte Phänomen ist auf dem Bild unterstrichen. Die Markierung, die sich auf das besprochene Phänomen bezieht, steht in Fettschrift.

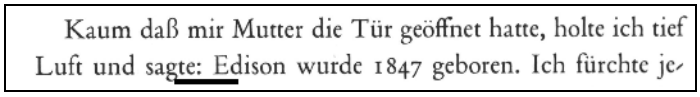
```
  Sollten Sie es sich aber doch noch überlegen …
</s>
```

e) ein Satz beginnt mit einem Großbuchstaben und endet mit einem Doppelpunkt. Dieses Zeichen kommt als Satzende nur am Absatzende vor:


Was sein würde, das wußte ich zufällig ganz genau:

```
<s xml:id="div56.p786.seg2451.s2572_jku00de">
  Was sein würde, das wußte ich zufällig ganz genau:
</s>
```

- in anderen Fällen bezeichnet ein Doppelpunkt die Grenze zwischen der indirekten und direkten Rede. Eine derartige Konstruktion wird hier als ein Satz kodiert:

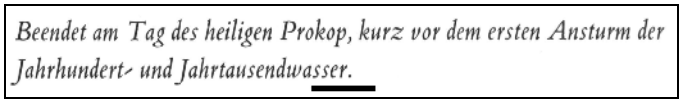

Kaum daß mir Mutter die Tür geöffnet hatte, holte ich tief Luft und sagte: Edison wurde 1847 geboren. Ich fürchte je-

```
<s xml:id="div4.p68.seg206.s218_jku00de">
  Kaum daß mir Mutter die Tür geöffnet hatte, holte ich tief
  <lb n="33:26"/>Luft und sagte: Edison wurde 1847 geboren.
</s>
```

f) weitere Zeichen (auch Kodierungen), die am Satzende vorkommen können und die automatische Identifizierung beeinflussen:

f-1) ein Endtag der Hervorhebung – formatierter Text:

- Punkt und Endtag der Hervorhebung:


Beendet am Tag des heiligen Prokop, kurz vor dem ersten Ansturm der Jahrhundert- und Jahrtausendwasser.

```
<s xml:id="div66.p971.seg2991.s3133_jku00de">
  <hi rend="italic">
    Beendet am Tag des heiligen Prokop, kurz vor dem ersten
Ansturm der
    <lb n="296:32"/>Jahrhundert- und Jahrtausendwasser.
  </hi>
</s>
```

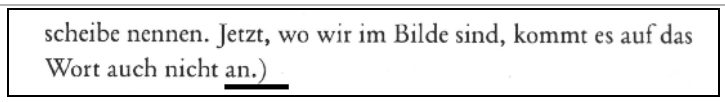
- Ausrufezeichen und Endtag der Hervorhebung:


Vstávej, Soňo, ty spáči!

```
<s xml:id="div57.p792.seg2466.s2587_jku00de">
  <hi rend="italic">Vstávej, Soňo, ty spáči!</hi>
</s>
```

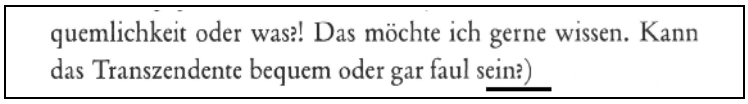
f-2) schließende Klammer:

- Punkt und schließende Klammer:


scheibe nennen. Jetzt, wo wir im Bilde sind, kommt es auf das Wort auch nicht an.)

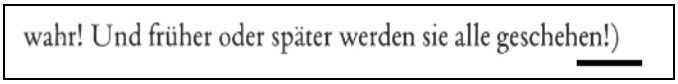
```
<s xml:id="div53.p683.seg2146.s2248_jku00de">
  Jetzt, wo wir im Bilde sind, kommt es auf das
  <lb n="218:7"/>Wort auch nicht an.)
</s>
```

- Fragezeichen und schließende Klammer:


quemlichkeit oder was?! Das möchte ich gerne wissen. Kann das Transzendente bequem oder gar faul sein?)

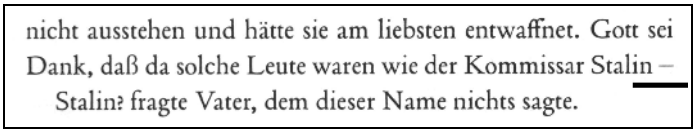
```
<s xml:id="div53.p677.seg2125.s2227_jku00de">Kann
  <lb n="216:24"/>das Transzendente bequem oder gar faul sein?)
</s>
```

- Ausrufezeichen und schließende Klammer:


wahr! Und früher oder später werden sie alle geschehen!)

```
<s xml:id="div32.p419.seg1308.s1380_jku00de">
  Und früher oder später werden sie alle geschehen!)
</s>
```

f-3) Geviertstrich (als Satzgrenze nur am Absatzende):


nicht ausstehen und hätte sie am liebsten entwaffnet. Gott sei Dank, daß da solche Leute waren wie der Kommissar Stalin – Stalin? fragte Vater, dem dieser Name nichts sagte.

```
<s xml:id="div25.p281.seg928.s972_jku00de">Gott sei
  <lb n="105:27"/>
  Dank, daß da solche Leute waren wie der Kommissar Stalin —
</s>
```

Die Adressierung mit dem Attribut `xml:id` wird durch die Angabe des `s`-Wertes erweitert, der die Nummerierung der Sätze vom Anfang des Romans an ergänzt.

In diesem Projekt war es aus technischen Gründen nicht möglich, die erste Auflage des Ausgangstextes zu verwenden, deswegen kam es zu der Sondersituation,

dass die deutsche Übersetzung vier zusätzliche Sätze im Vergleich zu dem hier benutzten Ausgangstext enthält. Damit die Regeln der Synoptisierung nicht gestört werden, werden leere `<s>`-Elemente (bzw. auch `<seg>`-Elemente) in den tschechischen Text eingefügt. Diese Elemente enthalten zusätzlich noch das Attribut `rend`, das das Element und seine Abbildung im Quelltext näher beschreibt. In diesem Fall ist der Wert `missing`, das bedeutet, dass das Phänomen im Quelltext nicht vertreten ist.

Ausgangssprache:

Obávám se, že tak malá zas není, upozornila jsem ho. A ostatně ani lidi nejsou tak milí.

```
<seg xml:id="div50.p658.seg2040_jkn05cz">
<s xml:id="div50.p658.seg2040.s2038_jkn05cz">Obávám se, že tak
malá zas není, upozornila jsem ho.
</s></seg>
<seg xml:id="div50.p658.seg2041_jkn05cz" rend="missing">
<s xml:id="div50.p658.seg2041.s2039_jkn05cz" rend="missing"/>
</seg>
<seg xml:id="div50.p658.seg2042_jkn05cz">
<s xml:id="div50.p658.seg2042.s2040_jkn05cz">A ostatně
<lb n="153:36"/>ani lidi nejsou tak milí.
</s></seg></p>
```

Zielsprache:

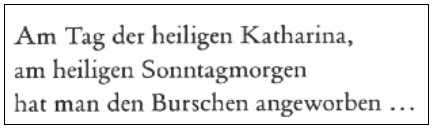
Ich fürchte, so klein ist es nun auch wieder nicht, wandte ich ein. Und im übrigen bin auch ich nicht mehr die Kleine. Und auch die Menschen hier sind nicht so nett.

```
<seg xml:id="div50.p658.seg2040_jku00de">
<s xml:id="div50.p658.seg2040.s2140_jku00de">Ich fürchte, so
klein ist es nun auch wieder nicht, wandte ich
<lb n="210:12"/>ein.
</s></seg>
<seg xml:id="div50.p658.seg2041_jku00de">
<s xml:id="div50.p658.seg2041.s2141_jku00de">Und im übrigen bin
auch ich nicht mehr die Kleine.
</s></seg>
<seg xml:id="div50.p658.seg2042_jku00de"><s
xml:id="div50.p658.seg2042.s2142_jku00de">Und
<lb n="210:13"/>auch die Menschen hier sind nicht so
nett.</s></seg>
```

Alle Text-Stellen mit fehlenden Sätzen sind im Header des XML-Dokuments verzeichnet.

3.2.5 `<lg>`

→ ‚line group' / Zeilengruppe – innerhalb von `<lg> … </lg>` ist das Gedicht in einzeln kodierten Verszeilen enthalten.

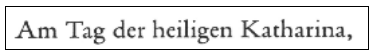



```
<lg xml:id="div46.lg583_jku00de">
  <l xml:id="div46.lg583.l22_jku00de">
    <seg xml:id="div46.p583.seg1826_jku00de">
      Am Tag der heiligen Katharina,
    </seg>
  </l>
  <lb n="188:12"/>
  <l xml:id="div46.lg583.l23_jku00de">
    <seg xml:id="div46.p583.seg1827_jku00de">
      am heiligen Sonntagmorgen
    </seg>
  </l>
  <lb n="188:13"/>
  <l xml:id="div46.lg583.l24_jku00de">
    <seg xml:id="div46.p583.seg1828_jku00de">
      hat man den Burschen angeworben …
    </seg>
  </l>
</lg>
```

Das Element enthält das Attribut `xml:id`, die Adressierung wird um die Angabe `lg` erweitert. Die Zeilengruppe (`lg`) ist auf dem gleichen Niveau wie der Absatz (Paragraph), statt `p` wird die Bezeichnung `lg` in der Adressierung verwendet und die Nummerierung bei `lg` setzt sich dort fort, wo sie bei dem vorhergehenden Absatz (`p`) aufgehört hat.

3.2.6 `<l>`

→ ‚verse line' / Verszeile – eine Verszeile umklammern die Tags `<l> … </l>`.




```
<l xml:id="div46.lg583.l22_jku00de">
  <seg xml:id="div46.p583.seg1829_jku00de">
    Am Tag der heiligen Katharina,
  </seg>
</l>
```

Das Element enthält das Attribut `xml:id`, die Adressierung wird um die Angabe `l` erweitert. Die Verszeile ist auf dem gleichen Niveau wie der Satz bzw. die Satzeinheit

(‚s-Unit'), deswegen wird statt `s` die Markierung `l` in der Adressierung verwendet. Die Nummerierung bei `l` (Verszeile) zählt nur die Verszeilen, ohne dass sie an die Satznummerierung anknüpfen würde.

3.2.7 <opener>

→ die Tags `<opener> … </opener>` bilden einen Container für die Grußformel am Anfang eines Briefes (dieses Element kommt nur im Kratochvil-Text vor).

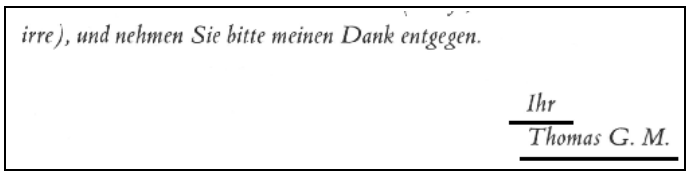

Teuerster Freund!
Ich schreibe nur noch schlecht und mit Mühe, aber gerade deshalb ist es

```
<opener>
  <salute xml:id="div33.sal424_jku00de">
    <seg xml:id="div33.p424.seg1327_jku00de">
      <hi rend="italic">Teuerster Freund!</hi>
    </seg>
  </salute>
</opener>
```

In diesem Element werden keine Attribute verwendet.

3.2.8 <closer>

→ die Tags `<closer> … </closer>` bilden einen Container für die Abschiedsformel am Ende eines Briefes (dieses Element kommt nur im Kratochvil-Text vor).


irre), und nehmen Sie bitte meinen Dank entgegen.

Ihr
Thomas G. M.

```
<closer>
  <salute xml:id="div33.sal426_jku00de">
    <seg xml:id="div33.p426.seg1351_jku00de">
      <hi rend="italic">Ihr
      <lb n="142:29"/>Thomas G. M.</hi>
    </seg>
  </salute>
</closer>
```

In diesem Element werden keine Attribute verwendet. Die Information zu der Position der Abschlussformel und der Unterschrift (dass sie im Text rechtsbündig steht) wird im Dokument-Header festgehalten.

3.2.9 <salute>

→ ‚salutation' / Anrede, Begrüßung – durch die Tags <salute> … </salute> wird eine Grußformel kodiert, wie in einem Brief. Ein Brief bzw. das Element salute kommt nur im Kratochvil-Text vor.

Die Grußformel am Briefanfang:

Teuerster Freund!

```
<opener>
  <salute xml:id="div33.sal424_jku00de">
    <seg xml:id="div33.p424.seg1327_jku00de">
      <hi rend="italic">Teuerster Freund!</hi>
    </seg>
  </salute>
</opener>
```

Die Grußformel am Briefende:

Ihr
Thomas G. M.

```
<closer>
  <salute xml:id="div33.sal426_jku00de">
    <seg xml:id="div33.p426.seg1351_jku00de">
      <hi rend="italic">Ihr
      <lb n="142:29"/>Thomas G. M.</hi>
    </seg>
  </salute>
</closer>
```

Es werden keine Attribute verwendet, um die Anfangs- und Abschiedsgrußformel zu unterscheiden, dies wird bereits durch ihre Platzierung im entsprechenden Container signalisiert (s. <opener> und <closer>).

Das <salute>-Element enthält das Attribut xml:id. Die Begrüßung ist auf dem gleichen Niveau wie der Absatz (‚Paragraph'), statt p wird die Bezeichnung sal in der Adressierung verwendet und die die Nummerierung bei sal setzt sich also dort fort, wo sie bei dem bevorstehenden Absatz (p) aufgehört hat.

3.2.10 <seg>

→ ‚arbitrary segment' / ein beliebiges Segment – mit den Tags <seg> … </seg> wird ein Textabschnitt für bestimmte zusätzliche Zwecke gekennzeichnet, die die Hierarchie der Textelemente ergänzen. Dieses Element wird hier als Hilfsstruktur für die Parallelisierung (Synoptisierung) der Sätze im Korpus verwendet. In der Regel steht

innerhalb eines Segments jeweils ein Satz[51] der Ausgangs- und der Zielsprache. Auf diese Weise kann ein Satz der Ausgangssprache einem Satz der Zielsprache bei der Parallelisierung der im DeuCze-Korpus angezeigten Texte entgegengestellt werden, z. B. dem Segment Nr. 26 in der Ausgangssprache entspricht das Segment Nr. 26 in der Zielsprache:

Ausgangssprache:

```
<seg xml:id="div1.p8.seg26_jkn05cz">
  <s xml:id="div1.p8.seg26.s28_jkn05cz">
    Ale pak se náhle stalo něco zvláštního.
  </s>
</seg>
```

Zielsprache:

```
<seg xml:id="div1.p8.seg26_jku00de">
  <s xml:id="div1.p8.seg26.s26_jku00de">
    Aber dann geschah plötzlich etwas Merkwürdiges.
  </s>
</seg>
```

Auch einzelne Verszeilen befinden sich jeweils in einem eigenen Segment[52]:

Ausgangssprache:

```
<l xml:id="div18.lg163.l6_jkn05cz">
  <seg xml:id="div18.p163.seg535_jkn05cz">Císaře a naši zem …
  </seg>
</l>
```

Zielsprache:

```
<l xml:id="div18.lg163.l6_jku00de">
  <seg xml:id="div18.p163.seg535_jku00de">Unsern guten Kaiser
Franz …
  </seg>
</l>
```

Ebenso werden auch die Grußformeln in jeweils einem Segment platziert:

Ausgangssprache:

```
<salute xml:id="div33.sal424_jkn05cz">
  <seg xml:id="div33.p424.seg1327_jkn05cz">
```

51 Nicht nur ein Satz, sondern jeweils eine satzartige Texteinheit, wie hier die Verszeile und die Grußformel.

52 Obwohl eine Verszeile auf dem gleichen Niveau wie ein Satz ist, ist die Reihenfolge der Elemente umgekehrt (Segment innerhalb von Verszeile: `<l><seg>…</seg></l>`, nicht wie bei einem Satz: `<seg><s>…</s></seg>`), sonst wäre das Dokument nicht valide. Ähnlich ist es auch bei ‚Salute'.

```
    <hi rend="italic">Drahý příteli!</hi>
  </seg>
</salute>
```

Zielsprache:

```
<salute xml:id="div33.sal424_jku00de">
  <seg xml:id="div33.p424.seg1327_jku00de">
    <hi rend="italic">Teuerster Freund!</hi>
  </seg>
</salute>
```

Eine Voraussetzung für die Synoptisierung der verwendeten Texte ist die Tatsache, dass der Ausgangstext und der Zieltext dieselbe Anzahl der Segmente bzw. Texteinheiten haben (vgl. DIAS 2005, 257f). Im Idealfall steht im Ausgangstext in einem Segment nur ein Satz und im Zieltext (d. h. im Segment mit Übersetzung) befindet sich ebenfalls nur ein Satz.

Die Gesamtsumme der Segmente im Kratochvil-Text ist 2 991, aus dieser Menge sind es 180 Segmente (d. h. 6 %), die ein anderes Übersetzungs-Verhältnis als 1:1 repräsentieren, es sind 28 Segmente in der tschechischen und 152 in der deutschen Datei. Die Segmente selbst stehen bei der Synoptisierung alle im Verhältnis 1:1, egal, ob sie einen oder mehrere Sätze enthalten. Mit Probe-Synoptisierungen wird geprüft, ob die Segmentgrenzen im Text richtig eingesetzt sind, damit sie in den verglichenen Dateien die entsprechenden Textabschnitte enthalten. Die Anpassung der Segmentgrenzen muss manuell durchgeführt werden. In das Bearbeitungsskript müssen die betroffenen Stellen eine nach der anderen eingetragen werden, damit sie bei der automatischen Bearbeitung entsprechend behandelt werden. Folgende Tabelle bietet eine Übersicht der gefundenen Übersetzungs-Verhältnisse im Kratochvil-Text, die anders als 1:1 sind:

der tschechische Text		der deutsche Text	
Übersetzungs-Verhältnis	Anzahl der betr. Segmente	Übersetzungs-Verhältnis	Anzahl der betr. Segmente
- -	- -	de 5 : 1 cz	1
cz 2 : 4 de	1	de 4 : 2 cz	1
- -	- -	de 4 : 1 cz	5
cz 2 : 3 de	1	de 3 : 2 cz	1
cz 3 : 1 de	1	de 3 : 1 cz	17
cz 2 : 2 de	2	de 2 : 2 cz	2
cz 2 : 1 de	23	de 2 : 1 cz	125
Σ	**28**	**Σ**	**152**

Tabelle 5: Übersetzungsverhältnis beim Kratochvil-Text

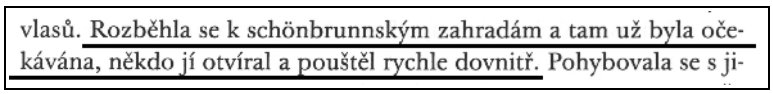
In diesem Text ist die Übersetzungsrichtung vom Tschechischen ins Deutsche. Die Tabelle wird danach aufgeteilt, in welchem Text die jeweilige Anpassung der Segmentgrenze durchgeführt wurde. Es wurden insgesamt 28 Eingriffe im tschechischen Text und 152 Eingriffe im deutschen Text in die XML-Struktur zusätzlich durchgeführt.

Es können jedoch weitere interessante Angaben ausgelesen werden, wie z. B. dass bei 148 Segmenten ein tschechischer Satz[53] mit mehreren deutschen Sätzen übersetzt wurde, dabei entsprechen in 125 Fällen zwei deutsche Sätze jeweils einem Satz aus dem tschechischen Originaltext. Nur in 24 Fällen wurden mehrere tschechische Sätze jeweils als ein deutscher Satz übersetzt, z. B. entsprechen 23 Mal zwei tschechische Sätze einem Satz im Zieltext und einmal werden drei tschechische Sätze als ein deutscher Satz übersetzt. Das Verhältnis 2:2 bedeutet zwar die gleiche Anzahl der enthaltenen Sätze im Segment, doch ihre Grenze ist in der Übersetzung im Vergleich zum Ausgangstext verschoben. Weiter unten folgen konkrete Beispiele.

Ein Beispiel für einen tschechischen Satz, der bei seiner Übersetzung ins Deutsche in zwei Sätze getrennt wurde:

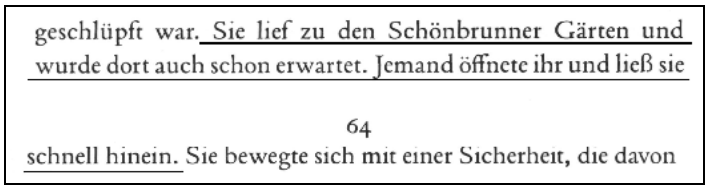
53 Gemeint ist ein Satz, wie es im Unterkapitel zum <s>-Element beschrieben wurde, nämlich ein Satz im typografischen Sinne.

Ausgangssprache:


vlasů. Rozběhla se k schönbrunnským zahradám a tam už byla očekávána, někdo jí otvíral a pouštěl rychle dovnitř. Pohybovala se s ji-

```
<seg xml:id="div17.p150.seg502_jkn05cz">
  <s xml:id="div17.p150.seg502.s502_jkn05cz">
    Rozběhla se k schönbrunnským zahradám a tam už byla oče-
    <lb n="50:29"/>kávána, někdo jí otvíral a pouštěl rychle
dovnitř.
  </s>
</seg>
```

Zielsprache:


geschlüpft war. Sie lief zu den Schönbrunner Gärten und wurde dort auch schon erwartet. Jemand öffnete ihr und ließ sie

64

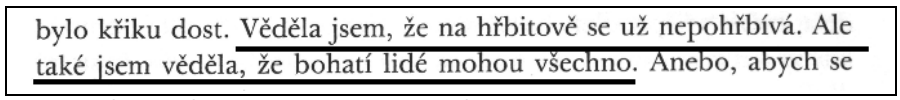
schnell hinein. Sie bewegte sich mit einer Sicherheit, die davon

```
<seg xml:id="div17.p150.seg502_jku00de">
  <s xml:id="div17.p150.seg502.s529_jku00de">
    Sie lief zu den Schönbrunner Gärten und
    <lb n="64:31"/>wurde dort auch schon erwartet.
  </s>
  <s xml:id="div17.p150.seg502.s530_jku00de">
    Jemand öffnete ihr und ließ sie
    <pb n="65"/><lb n="65:1"/>schnell hinein.
  </s>
</seg>
```

Das Beispiel zeigt, dass der tschechische Originalsatz mit zwei deutschen Sätzen übersetzt wird, und dass also das deutsche Segment zwei Sätze enthält. Die Zugehörigkeit dieser Textpassagen zueinander ist auch an der gleichen Nummer des Segmentes zu erkennen (hier 502).

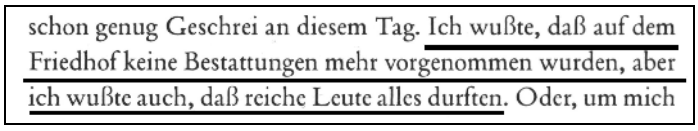
In manchen Fällen werden zwei oder mehrere Sätze bei der Übersetzung verbunden, in diesem Beispiel werden zwei tschechische Sätze als ein deutscher Satz übersetzt:

Ausgangssprache:


bylo křiku dost. Věděla jsem, že na hřbitově se už nepohřbívá. Ale také jsem věděla, že bohatí lidé mohou všechno. Anebo, abych se

```
<seg xml:id="div23.p229.seg772_jkn05cz">
  <s xml:id="div23.p229.seg772.s774_jkn05cz">
    Věděla jsem, že na hřbitově se už nepohřbívá.
  </s><s xml:id="div23.p229.seg772.s775_jkn05cz">
    Ale<lb n="70:14"/>také jsem věděla, že bohatí lidé mohou
všechno.</s></seg>
```

Zielsprache:

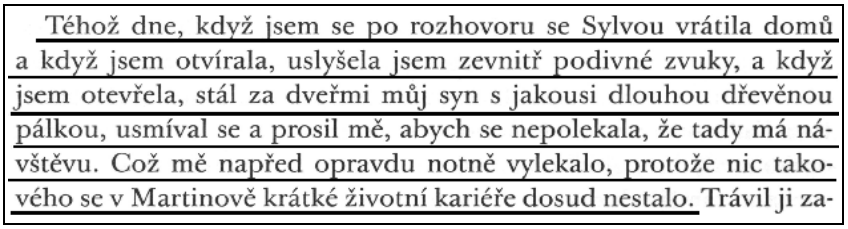

schon genug Geschrei an diesem Tag. Ich wußte, daß auf dem Friedhof keine Bestattungen mehr vorgenommen wurden, aber ich wußte auch, daß reiche Leute alles durften. Oder, um mich

```
<seg xml:id="div23.p229.seg772_jku00de">
  <s xml:id="div23.p229.seg772.s809_jku00de">
    Ich wußte, daß auf dem
    <lb n="90:18"/>Friedhof keine Bestattungen mehr vorgenommen
wurden, aber
    <lb n="90:19"/>ich wußte auch, daß reiche Leute alles
durften.
  </s>
</seg>
```

Wie die einleitende tabellarische Übersicht gezeigt hat, können im Text auch andere Übersetzungsverhältnisse vorkommen. Die Anzahl der Sätze bei der Ausgangs- und Zielsprache schwankt im hier behandelten Text bis zu fünf Sätzen, die jeweils sowohl nur einem Satz als auch mehreren Sätzen in der entgegengesetzten Sprache entsprechen können. Gemeint sind damit auch Fälle, wo z. B. zwei tschechische Sätze als drei deutsche Sätze übersetzt werden und so ähnlich. Aus Platzgründen werden hier nicht alle Belege für diese weiteren unterschiedlichen Verhältnisse abgebildet, das Prinzip der Anpassung der Segmentgrenze ist dann ähnlich wie bei den angeführten Beispielen.

Trotzdem sei noch mit einem Beispiel auf die interessanten Fälle eingegangen, wo die Anzahl der Sätze im Segment gleich ist und die Satzgrenze bei der Übersetzung verschoben wurde:

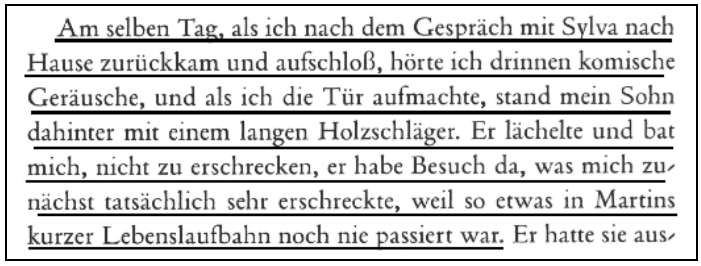
Ausgangssprache:


Téhož dne, když jsem se po rozhovoru se Sylvou vrátila domů a když jsem otvírala, uslyšela jsem zevnitř podivné zvuky, a když jsem otevřela, stál za dveřmi můj syn s jakousi dlouhou dřevěnou pálkou, usmíval se a prosil mě, abych se nepolekala, že tady má návštěvu. Což mě napřed opravdu notně vylekalo, protože nic takového se v Martinově krátké životní kariéře dosud nestalo. Trávil ji za-

```
<seg xml:id="div50.p646.seg2014_jkn05cz">
  <s xml:id="div50.p646.seg2014.s2011_jkn05cz">
    Téhož dne, když jsem se po rozhovoru se Sylvou vrátila domů
    <lb n="152:23"/>a když jsem otvírala, uslyšela jsem zevnitř
      podivné zvuky, a když
    <lb n="152:24"/>jsem otevřela, stál za dveřmi můj syn s
jakousi dlouhou dřevěnou
    <lb n="152:25"/>pálkou, usmíval se a prosil mě, abych se
      nepolekala, že tady má ná-
```

```
    <lb n="152:26"/>vštěvu.
  </s>
  <s xml:id="div50.p646.seg2014.s2012_jkn05cz">
    Což mě napřed opravdu notně vylekalo, protože nic tako-
    <lb n="152:27"/>vého se v Martinově krátké životní kariéře
dosud nestalo.
  </s></seg>
```

Zielsprache:



Am selben Tag, als ich nach dem Gespräch mit Sylva nach Hause zurückkam und aufschloß, hörte ich drinnen komische Geräusche, und als ich die Tür aufmachte, stand mein Sohn dahinter mit einem langen Holzschläger. Er lächelte und bat mich, nicht zu erschrecken, er habe Besuch da, was mich zu-nächst tatsächlich sehr erschreckte, weil so etwas in Martins kurzer Lebenslaufbahn noch nie passiert war. Er hatte sie aus-

```
<seg xml:id="div50.p646.seg2014_jku00de">
  <s xml:id="div50.p646.seg2014.s2113_jku00de">
    Am selben Tag, als ich nach dem Gespräch mit Sylva nach
    <lb n="208:16"/>Hause zurückkam und aufschloß, hörte ich
drinnen    komische
    <lb n="208:17"/>Geräusche, und als ich die Tür aufmachte,
stand mein  Sohn
    <lb n="208:18"/>dahinter mit einem langen Holzschläger.
  </s>
  <s xml:id="div50.p646.seg2014.s2114_jku00de">Er lächelte und
bat
    <lb n="208:19"/>mich, nicht zu erschrecken, er habe Besuch
da, was    mich zu-
    <lb n="208:20"/>nächst tatsächlich sehr erschreckte, weil so
etwas in   Martins
    <lb n="208:21"/>kurzer Lebenslaufbahn noch nie passiert war.
  </s>
</seg>
```

Auch das `<seg>`-Element wird jeweils mit dem Attribut `xml:id` versehen. Der Wert dieses Attributs wird um die Angabe `seg` erweitert, die die Nummerierung der Segmente im Text angibt.

Es wäre sicherlich interessant zu untersuchen, ob es gewisse Regelmäßigkeiten und Gründe dafür gibt, wie bestimmte Sätze bei ihrer Übersetzung geteilt oder verknüpft werden. Die erstellte XML-Datei legt eine Grundlage für solche Untersuchungen, denn die Segmente, die mehrere Sätze enthalten, wurden bei der Parallelisierung der Texte entsprechend attribuiert. Diese Attribute sind in dieser Projektphase zwar ausgeblendet, bei Bedarf können sie jedoch durch das PHP-Bearbeitungsskript in

die Zieldatei eingeführt werden. Eine Analyse der unterschiedlichen Übersetzungsverhältnisse wurde nicht durchgeführt, dies würde den Rahmen dieser Arbeit sprengen.

Nachdem alle Segmentgrenzen entsprechend angepasst werden, ist die parallele Ausgabe dieser Texteinheiten im Verhältnis eins zu eins auf der Benutzeroberfläche des Korpus möglich.

Das Synoptisierungsskript an sich wird hier nicht beschrieben, da es von anderen Projekt-Mitgliedern erstellt wurde.

3.2.11 `<pb/>`

→ ‚page break' / Seitenumbruch – steht am Anfang der jeweiligen Seite.

```
<pb n="7"/>
```

Das enthaltene Attribut `n` (‚number') gibt die Seitennummer an.

3.2.12 `<lb/>`

→ ‚line break' / Zeilenumbruch – steht am Anfang der jeweiligen Zeile.

```
<lb n="7:2"/>
```

Dieses Element enthält das `n`-Attribut mit einem Wert, der die Seiten- (links vom Doppelpunkt) und die Zeilennummer (rechts vom Doppelpunkt) angibt.

3.2.13 `<hi>`

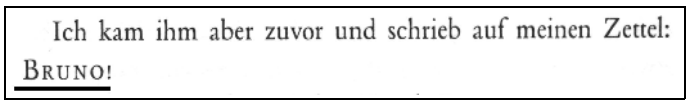
→ ‚highlighted' / Hervorhebung. Der hervorgehobene Text wird durch `<hi>` … `</hi>` geklammert. Durch den Wert des Attributs `rend` wird die Art der Hervorhebung angegeben: `italic` für Kursive, `smallCaps` für Kapitälchen und `bold` für Fettschrift:

- Kursivschrift:

und zwei aufgeschlagene Zeitungen: *Moskowskije nowosti* und *Frankfurter Allgemeine Zeitung.*

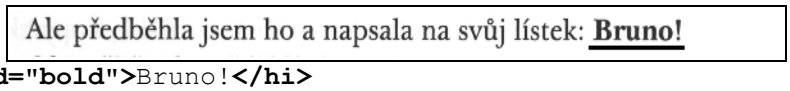
```
<hi rend="italic">Frankfurter Allgemeine Zeitung</hi>
```

- Kapitälchenschrift:


Ich kam ihm aber zuvor und schrieb auf meinen Zettel:
BRUNO!

```
<hi rend="smallCaps">Bruno!</hi>
```

- Fettschrift[54]:

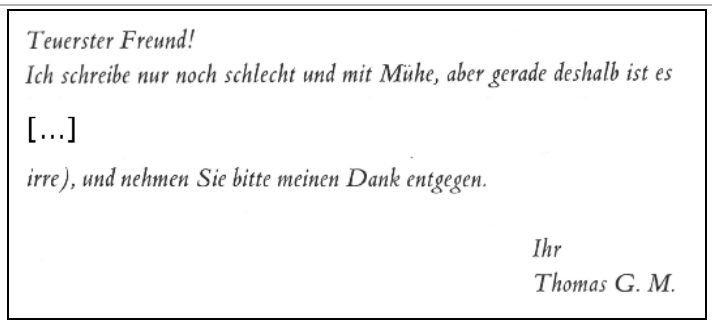

Ale předběhla jsem ho a napsala na svůj lístek: **Bruno!**

```
<hi rend="bold">Bruno!</hi>
```

Sobald die Hervorhebung die Satzgrenze überschreitet, muss sie vor dem Satzende-Tag geschlossen und nach dem nächsten Satzanfang-Tag wieder aufgenommen werden – auf diese Weise bleibt die Verschachtelung der Elemente erhalten und das Dokument ist immer noch wohlgeformt.

Im Kratochvil-Text betrifft dies die Unterbrechung der Kursivpassagen:

a) im Brief, wo die Grußformeln von dem Brieftext getrennt kodiert sind


Teuerster Freund!
Ich schreibe nur noch schlecht und mit Mühe, aber gerade deshalb ist es

[…]

irre), und nehmen Sie bitte meinen Dank entgegen.

Ihr
Thomas G. M.

```
<div>
  <opener>
    <salute xml:id="div33.sal424_jku00de">
      <seg xml:id="div33.p424.seg1327_jku00de">
        <hi rend="italic">Teuerster Freund!</hi>
      </seg>
    </salute>
  </opener>
  <lb n="141:23"/>
  <p xml:id="div33.p425_jku00de">
    <hi rend="italic">
      <seg xml:id="div33.p425.seg1328_jku00de">
        <s xml:id="div33.p425.seg1328.s1400_jku00de">Ich schreibe
nur noch
            schlecht und mit Mühe, aber gerade deshalb ist es

        […]

        <lb n="142:27"/>irre), und nehmen Sie bitte meinen Dank
entgegen.
```

54 Das Beispiel für die Kodierung der Fettschrift stammt aus dem tschechischen Text von Jiří Kratochvil.

```
        </s>
      </seg>
    </hi>
  </p>
  <lb n="142:28"/>
  <closer>
    <salute xml:id="div33.sal426_jku00de">
      <seg xml:id="div33.p426.seg1351_jku00de">
        <hi rend="italic">Ihr
        <lb n="142:29"/>Thomas G. M.</hi>
      </seg>
      </salute>
  </closer>
</div>
```

b) bei den einzelnen Teilen der Gedichte

Robert Lowell
Worte für ein Meerschweinchen, genannt Buchtel

```
<lg xml:id="div61.lg845_jku00de">
  <l xml:id="div61.lg845.l25_jku00de">
    <seg xml:id="div61.p845.seg2632_jku00de">
      <hi rend="italic">
        Robert Lowell,
      </hi>
    </seg>
  </l>
  <lb n="263:25"/>
  <l xml:id="div61.lg845.l26_jku00de">
    <seg xml:id="div61.p845.seg2633_jku00de">
      <hi rend="italic">
        Worte für ein Meerschweinchen, genannt Buchtel
      </hi>
    </seg>
  </l>
</lg>
```

Diese Beschreibung enthält alle Elemente und ihre Anwendung in dieser Phase des Projektes. In Zukunft ist es nicht ausgeschlossen, dass neue Elemente und Attribute die bestehende Anzahl erweitern werden. Anschließend werden die Ausgangstexte in eine XML-Datei transformiert, damit die satzweise synoptisierte Anzeige der bearbeiteten Texte im DeuCze-Korpus möglich ist.

3.3 Die Transformation von der Textdatei in XML

3.3.1 Vorbereitung der Texte

In dieser Phase stehen die gewonnenen Texte korrigiert zur Verfügung und entsprechen der Druckvorlage. Dank der Speicheroption ‚Formatierter Text' von ABBYY FineReader enthalten die im MS Word-Format gespeicherten Dateien die Textformatierung, so wie sie in der Druckvorlage vertreten ist. Bevor die Daten als eine Textdatei ohne Formatierung gespeichert werden, werden sie noch mit MS Word bearbeitet. Die formatierten Textstellen (wie Fett- und Kursivschrift) werden mit XML-Tags kodiert, damit die Information über diese Information beim Speichern in das Textformat nicht verloren geht.

Die ersten XML-Tags werden in folgenden Schritten eingesetzt:

a) der formatierte Text wird gesucht, um den gefundenen Text (bzw. die Zeichenkette) mit entsprechenden Tags zu versehen. Hierfür wird die eingebaute Suchfunktion von MS Word (‚Suchen und Ersetzen') verwendet. Die Ergänzung der Tags geschieht in folgenden Schritten:

- das Feld ‚Suchen nach' bleibt leer. In den erweiterten Einstellungen ‚Erweitern' wird unter ‚Format' die Suche nach Zeichen in ‚Kursiv' gewählt.

- im Feld ‚Ersetzen durch' wird angegeben, wie der gefundene Text zu behandeln ist: in den erweiterten Optionen wird ‚Sonstiges' und das Symbol für ‚Suchen nach Text' gewählt. Im Feld ‚Ersetzen durch' erscheint die Entität für den gesuchten Text (`^&`), zu der die gewünschten Tags zusätzlich eingetippt werden, hier z. B. die Kursivschrift: `<hi rend="italic">^&</hi>`

- das Ersetzen der betroffenen Stellen wird durch das Betätigen des Buttons ‚Ersetzen' (jeweils für eine Stelle) oder ‚Alle ersetzen' (für das ganze Dokument) durchgeführt

b) wenn nötig, wird eine Korrektur durchgeführt:

- die Korrekturen bestehen vor allem darin, dass die Reihenfolge bestimmter Zeichen ausgebessert werden muss. Z. B. wenn durch das automatische Ersetzen Formatierungsabschnitte ohne Text (wie: `<hi rend="italic"></hi>`) entstehen, dann müssen sie gelöscht werden.

Das Prinzip für das Markieren der Fett- und Kapitälchenschrift ist gleich, es müssen nur die entsprechenden XML-Tags in das Feld ‚Ersetzen durch' eingesetzt werden. Die Überschriften im deutschen Kratochvil-Text sind ebenfalls in Kapitälchenschrift gesetzt. Später wird mit dem PHP-Skript nach Zeilen gesucht, die nur Text in Kapitälchen enthalten, um diese als Überschriften (mit dem <head>-Tag) zu markieren. Durch visuelle Kontrolle wird überprüft, ob tatsächlich nur Überschriften markiert wurden, und die falsch markierten Stellen werden korrigiert.

Während der Bearbeitung in MS Word werden neben der Formatierung der betroffenen Textstellen auch die Absatzanfänge markiert. Die Absatzkennzeichnung im gedruckten Text wird durch den Einzug der ersten Zeile durchgeführt, aber z. B. im deutschen Kratochvil-Text beginnt der erste Absatz des jeweiligen Kapitels mit einer Initiale linksbündig, und im tschechischen Kratochvil-Text beginnt der erste Absatz des jeweiligen Kapitels ohne Initiale, aber auch linksbündig. Das OCR-Programm versuchte, die Absätze im Text mit einer Einrückung der ersten Zeile zu kennzeichnen, aber dies war leider nicht konsequent. Bei der automatischen Verarbeitung könnten diese Stellen mit Verszeilen, die alle eingerückt sind, kollidieren. Deswegen wird an den Anfang jedes Absatzes eine Art Hilfsmarkierung manuell eingesetzt, die nirgendwo anders im Text vorkommen darf, um sie als solche zu erkennen. Hier wurde die Zeichenfolge ‚XYQ' verwendet, die abschließend bei der Ergänzung der Absatz-Tags gelöscht wird.

Im letzten Schritt wird an den Anfang der Datei die Seitennummerierung eingefügt, damit die Nummerierung der Seiten mit der Ziffer ‚2' beginnt. Das Ziel dieser Zusatznummerierung ist es, dass in einer späteren Bearbeitungsphase die Zahlenreihe, die mit einer Zwei beginnt, automatisch als Seitennummerierung erkannt wird (im Unterschied zu der Kapitelnummerierung, die mit einer Eins beginnt), und die Zahlen werden später gegen die <pb/>-Tags ausgetauscht. Zum Schluss muss die Datei entsprechend gespeichert werden.

Das Speichern als Textdatei ohne Formatierung

Damit der Text ohne unerwünschte Formatierung gespeichert wird, wird er in eine Textdatei ohne Formatierung (sog. ‚plain text file') gespeichert. In MS Word wird im Speicherdialog im Fenster ‚Dateityp' der Typ ‚Nur Text' ausgewählt. Danach wird

der Dateiname eingegeben, und nach dem Drücken des ‚Speichern'-Buttons wird das Dialogfenster ‚Dateikonvertierung' angezeigt. Es wird der Radio-Button ‚Andere Codierung'[55] gewählt, und in der Liste wird ‚Unicode (UTF-8)' markiert. Nach der Bestätigung wird die gewünschte Datei (z. B. `quelltext_kratochvil_de.txt`) erstellt. Die Datei ist in dieser Phase für die nächste Bearbeitung vorbereitet, um in ein XML-Dokument umgewandelt zu werden.

3.3.2 Die eigentliche Transformation in das XML-Format

Soweit es im Folgenden nicht anders angegeben wird, werden die Schritte der Transformation der Klartextdatei in die gültige XML-Datei am Beispiel des deutschen Textes von Jiří Kratochvil beschrieben. Die Transformation in XML wurde unter dem Betriebssystem Ubuntu Linux durchgeführt.

Es wurde jeweils ein selbstständiges Skript bzw. eine Skript-Reihe für jeden Ausgangstext erzeugt (d. h. eins für den tschechischen und eins für den deutschen Text, dasselbe gilt für das Buch von Brussig). Obwohl die parallelen Skripte überwiegend dieselben Schritte durchführen, gibt es trotzdem viele Operationen, die die konkreten Textstellen nur in der einen gegebenen Datei behandeln. Besonders aus Gründen der Übersichtlichkeit und Wartung der Skripte ist es also günstiger, für jeden Text jeweils ein konkret angepasstes Skript zu verwenden.

Schematische Darstellung der Arbeitsschritte:

a) XML-Markierung der Elemente bis zur Satzebene

- die Ausgangsdatei: einfache Textdatei (ohne Formatierung), UTF-8[56]

  `quelltext_kratochvil_de.txt` bzw.
  `quelltext_kratochvil_cz.txt`

- das PHP-Skript – es fügt die XML-Struktur ein[57]

  `KR_plain_to_xml_kr_de.php` bzw.
  `KR_plain_to_xml_kr_cz.php`
  +
  `KR_plain_to_xml_funktionen.php`

- die Ausgabe: die Quelldatei ist in XML-Form

55 ‚Codierung' – die Schreibweise mit ‚c' wird dem MS Word entnommen; in dieser Arbeit wird sonst die Schreibweise mit ‚k' bevorzugt.

56 Die erste Seite des Romans in ihrer Quellform (nach der OCR-Bearbeitung und den folgenden Korrekturen) befindet sich im Anhang Nr. 4.

57 Das PHP-Skript wird im Anhang Nr. 5 angeführt, das Skript mit den Funktionen befindet sich im Anhang Nr. 6.

`zieldatei_KR_de_format_xml.xml` bzw.
`zieldatei_KR_cz_format_xml.xml`

b) die Nummerierung und Identifikationsattribute werden eingeführt

- die Ausgangsdatei dieser Phase ist die Ausgabe des vorhergehenden Schrittes
`zieldatei_KR_de_format_xml.xml` bzw.
`zieldatei_KR_cz_format_xml.xml`

- die XSLT-Schablone steuert die Zählung und das Ergänzen bestimmter Attribute und ihrer Werte
`KR_nummerieren.xsl`

- die Ausgabe: die XML-Datei enthält die fast komplette vorgesehene Kodierung (es fehlt nur noch der Header)
`zieldatei_KR_de_nummeriert.xml` bzw.
`zieldatei_KR_cz_nummeriert.xml`

c) der Header wird hinzugefügt

- die Ausgangsdatei ist die Ausgabe des vorhergehenden Schrittes
`zieldatei_KR_de_nummeriert.xml` bzw.
`zieldatei_KR_cz_nummeriert.xml`

- der Header wird erstellt und dann dem getaggten Romantext hinzugefügt
`header_krat_de.xml` bzw.
`header_krat_cz.xml`

- die komplette valide Zieldatei wird gespeichert
`krat_de.xml` bzw.
`krat_cz.xml`

Die einzelnen Bearbeitungsschritte werden im Folgenden näher beschrieben:

ad a) Als Ausgangsdatei dient die Textdatei, die durch die OCR-Bearbeitung der gescannten Bücher entstanden ist; der Zeichensatz ist Unicode UTF-8, das Layout (wie Zeilentrennung usw.) entspricht der Vorlage.

Unter Verwendung der PHP-Skripte wird die XML-Struktur mit Markierung bis zur Satzebene erzeugt. Für jeden Ausgangstext gibt es ein eigenes Skript, dies ist nicht nur wegen der jeweils verwendeten Sprache nötig, sondern auch wegen der unterschiedlichen typografischen Gestaltung der jeweiligen Texte. Zum Beispiel steht im deutschen Text die Nummerierung der Überschriften auf einer eigenen Zeile, und die

Überschrift selbst ist in Kapitälchenschrift gesetzt. Dagegen steht im tschechischen Text sowohl die Nummer als auch die Überschrift auf nur einer Zeile und ist in Normalschrift gesetzt. Später werden die Überschriften durch diese Charakteristik als Überschriften automatisch identifiziert. Wenn diese beiden Typen in einem einzigen Skript definiert wären, würde dies zur Unübersichtlichkeit der Wartung des Skriptes beitragen.

Damit die Datei nach den TEI P5-Regeln valide ist, steht der gesamte Datei-Inhalt innerhalb des `<TEI>`-Elements und das eigene Dokument an sich besteht aus den Teilen ‚Header' und ‚Text'. Der gesamte Romantext befindet sich im TEI-Element `<body>…</body>`. In dieser Phase wird der ‚Text'-Teil bearbeitet und der ‚Header' wird zusätzlich manuell eingefügt, damit bei der Tagnummerierung nicht die Tags des Headers mitgezählt werden.

Aus praktischen Gründen werden die verwendeten PHP-Funktionen[58] in eine selbstständige Datei ausgelagert. Auch ausgewählte PHP-Funktionen werden in Abhängigkeit davon geteilt, für welchen Ausgangstext sie verwendet werden.

Es sei hier daran erinnert, dass die nachstehende Beschreibung sich auf die Bearbeitung des deutschen Textes von Jiří Kratochvil (*Unsterbliche Geschichte*) bezieht. Es handelt sich konkret um die Dateien `KR_plain_to_xml_kr_de.php` (im Folgenden ‚PTX') und die Datei, aus der die PHP-Funktionen gelesen werden `KR_plain_to_xml_funktionen.php` (im Folgenden ‚F').

Es werden beinahe alle Textverarbeitungsschritte von Korrekturen begleitet, die nur die jeweils bearbeiteten Stellen betreffen. Deswegen werden hier nicht alle Skriptzeilen einzeln behandelt, sondern es werden jeweils nur die wesentlichen Schritte besprochen (konkrete Informationen zu einzelnen Operationen sind als Kommentar direkt im Skript zugänglich; bei jedem Schritt folgt ein Verweis auf die Datei und die behandelten Zeilen):

1) Die Variablen werden definiert – der Name der Quell- und Zieldatei wird eingegeben (PTX: 18 und 21).

58 Die PHP-Funktionen ermöglichen mehrere Prozesse zusammenzufassen und geben einen Wert zurück. Die Möglichkeit, dass sich die PHP-Funktionen in einer getrennten Datei befinden können, trägt wesentlich zur Übersichtlichkeit des Ausgangsskriptes bei.

2) Die Funktion `fragezeichen` (PTX: 34 und F: 3–154) wird durchgeführt. Zuerst wird die Datei zeilenweise eingelesen. Das Ziel ist es, die Markierung der ausgewählten Satzgrenzen, die durch Frage-, Ausrufezeichen und Auslassungspunkte gekennzeichnet sind, zu ergänzen. Diese Funktion ist für beide Ausgangssprachen gemeinsam, weil überwiegend mit Strukturen gearbeitet wird, die allgemein auf die Satzgrenze hindeuten, wie z. B. ‚ein Fragezeichen gefolgt von Großbuchstaben'.

   Das Hauptkriterium bei der Bestimmung der Satzgrenze geht davon aus, dass ein Satz mit einem konkreten Zeichen (Satzendepunkt, Frage- oder Ausrufezeichen, drei Punkte) endet und diese Stelle gleichzeitig den Anfang des Folgesatzes bedeutet. Satzgrenzen am Anfang und am Ende eines Absatzes werden durch die Absatzgrenzen identifiziert.

   Wenn nach dem Fragezeichen ein Großbuchstabe kommt, handelt es sich im Allgemeinen um eine Satzgrenze (in dem Sinne, dass das `<s>`- und `<seg>`-Tag eingesetzt wird), bzw. es können auch (öffnende oder schließende) Klammer oder Zeilenumbruch dazwischen vorkommen. Wenn nach einem Fragezeichen ein Komma, Ausrufezeichen, Gedankenstrich oder auch ein Zeilenumbruch gefolgt von einem Kleinbuchstaben vorkommt, wird keine Satzgrenze gesetzt. An solchen Stellen werden die Fragezeichen zusätzlich mit Unterstrichen versehen, als Markierung von ‚nicht-Satzende'. So wird die weitere Suche erleichtert. Diese provisorisch eingesetzten Zeichen werden abschließend nach der eindeutigen Markierung aller Fragezeichen entfernt.

   In ähnlicher Weise werden auch die Ausrufezeichen bearbeitet. Folgt nach einem Ausrufezeichen ein Kleinbuchstabe, dann wird an dieser Stelle keine Satzgrenze eingetragen. Sobald am Absatzende Auslassungspunkte stehen, werden auch diese gleich als Satzende markiert.

   Je nach dem Quelltext werden die Suchanfragen verändert, weil neue Zeichen hinzutreten können, wie z. B. unterschiedliche Arten von Anführungszeichen, Gedankenstrich oder auch die XML-Markierung für Kursiv-, Fett- oder Kapitälchenschrift.

   Diese Funktion (`fragezeichen`) gibt den gesamten Text in einer Variablen für weitere Bearbeitung in das Ausgangsskript (PTX: 34) zurück.

3) Überschriftenmarkierungen herstellen (PTX: 42–102):

   - die Überschriften in diesem Text bestehen jeweils aus einer Ziffer (Nummerierung) und Text. In dieser Bearbeitungsphase sind die Überschriften bereits mit den TEI-Tags für die Markierung von Kapitälchenschrift (`<hi rend="smallCaps"> ... </hi>`) versehen. Dies wurde in einer früheren Bearbeitungsphase mit MS Word durchgeführt (vgl. Unterkapitel „3.3.1 Vorbereitung der Texte“). Diese Tags werden in `<head>`-Tags umgewandelt. Die Zahlen sind in Normalschrift und stehen in eigenen Zeilen und sind in dieser Phase in der Überschriftmarkierung noch nicht einbezogen. Im Text gibt es jedoch auch solche Stellen, die zwar in Kapitälchenschrift gesetzt sind, aber keine Überschriften bezeichnen. Es folgt Kontrolle und Korrektur der falsch markierten Stellen (z. B. Textabschnitte, die in Kapitälchenschrift stehen sollen, ohne dass es sich um Überschriften handelt). Wenn alle Überschriften korrekt markiert sind, wird ihnen die Nummerierung zugeordnet (PTX: 88–102). Der satzweise segmentierte Inhalt wird wieder in einer Variablen zum Ganztext verbunden und es werden Korrekturen durchgeführt, wie z. B. das Löschen von überflüssigen Leerzeilen.

4) Markierung der Seitennummern (PTX: 131–148):

   Der Gesamttext wird in eine Variable zeilenweise eingelesen, um nach dem Inhalt der Zeilen die Seitennummern zu identifizieren. Der Inhalt der Zeilen, die nur eine Zahl ohne weiteren Text enthalten[59], wird gelöscht und an seine Stelle kommt das Tag `<pb/>` (‚page break‘, Seitenumbruch). Die Nummerierung wird später mit der XSLT-Schablone eingefügt.

5) Die Funktion `austauschen`, mit der die Markierung der Satzgrenzen abgeschlossen wird, wird aufgerufen (PTX: 152 und F: 156–253): Zuerst werden die mit der Hilfsmarkierung ‚XYQ‘ gekennzeichneten Absatzanfänge in die korrekten Absatz-Tags zusammen mit Satzgrenzen-Tags eingefügt. Das bedeutet, dass statt ‚XYQ‘ in den Text die Tag-Folge `</s></seg></p><p><seg><s>` kommt.

59 Dies wird so durchgeführt, dass es eigentlich nach einer mit der Ziffer Eins beginnenden Zahlenreihe gesucht wird und diese Zahlenreihe wird bearbeitet.

Anschließend werden die restlichen Punkt-Zeichen behandelt; oft hilft auch die unmittelbare Umgebung, die Satzgrenze zu identifizieren. Wenn z. B. die Auslassungspunkte am Absatzende vorkommen, dürfen sie automatisch als Satzende behandelt werden. Durch das Ausschlussverfahren werden alle Punkte, die nicht als Satzendepunkte fungieren, von den richtigen Satzendepunkten unterschieden. Alle Auslassungspunkte, Punkte nach Ordinalzahl oder nach einer römischen Zahl und Abkürzungspunkte wurden hiermit temporär durch `#` ersetzt. Es kann davon ausgegangen werden, dass die übrigen Punkte im Text die Satzgrenze bezeichnen; sie werden entsprechend als Satzgrenze markiert und danach die `#`-Zeichen wieder in das Punkt-Zeichen umgewandelt.
Einen untrennbaren Teil dieser Funktion stellen verschiedene Korrekturen und zusätzliche Anpassungen dar, besonders muss die falsche Reihenfolge der Tags ausgebessert werden. Die Funktion `austauschen` übergibt den bearbeiteten Text zur Behandlung in das Hauptskript zurück.

6) TEI – an den Anfang und an das Ende der Datei wird die TEI-Markierung eingefügt (PTX: 167 und 170); der eigentliche Buchtext befindet sich innerhalb dieser Elemente.

   Die folgenden drei Funktionen sind den beiden Ausgangstexten nicht mehr gemeinsam. Wegen der Suche nach konkreten Textstellen werden die Funktionen so aufgeteilt, dass jeweils eine für den deutschen und eine für den tschechischen Text zur Verfügung steht.

7) Die Funktion `zusatz_korrektur_de` (PTX: 178 und F: 255–412) – es handelt sich um spezifische Korrekturen, z. B. Markierung der Überschriften der Romanbücher. Die überflüssigen Zeilenumbrüche werden gelöscht, und andere Sonderfälle wie Reihenfolge falsch positionierter Tags werden ausgebessert.

8) Die Funktion `synop_de` (PTX: 180 und F: 554–847) – bei der Übersetzung kann es passieren, dass z. B. ein Satz mit zwei Sätzen übersetzt wird. In diesen Fällen ist es wichtig, dass die zwei Sätze der Zielsprache in ein einziges Segmentelement positioniert werden. Auf diese Weise können die Sätze für die parallele

zweisprachige Ausgabe der Korpustexte synoptisiert, d. h. im Verhältnis 1:1 nebeneinander dargestellt werden. Innerhalb dieser Funktion werden die überflüssigen Segmentgrenzen gelöscht, und deswegen ist diese Funktion streng an den jeweiligen Text gebunden. (Mehr zu dieser Problematik im Unterkapitel zur Beschreibung des `<seg>`-Elements „3.2.10 `<seg>`“.) Alle `<s>`- und `<seg>`-Elemente, bei denen die Satzgrenze angepasst wurde, werden mit dem Attribut `attr` markiert. Der Wert dieses Attributs gibt das Übersetzungsverhältnis an, das anders als 1:1 ist. Der Wert des Attributs `attr` ist z. B. `2_1`, wenn zwei Sätze als ein Satz übersetzt wurden, usw. Während der Standardbearbeitung werden jedoch diese zusätzlichen Attribute gelöscht und nicht in die Zieldatei übertragen. Wenn die Zeilen F: 775 (für den deutschen Text) und F: 845 (für den tschechischen Text) auskommentiert werden[60], können im Skript diese Attribute bei den Sätzen beibehalten werden.

9) Die Funktion `sonstiges_de` (PTX: 182 und F: 849–915) fasst solche Fälle zusammen, die nachträglich behandelt werden müssen, wie die Markierung der Gedichte oder des (nur im Kratochvil-Text vorkommenden) Briefs.
   - Markierung der Gedichte – der Absatz-Tag `<p>` wird in das Zeilengruppe-Tag umgewandelt (`<lg>`) und statt Satz-Tag (`<s>`) wird Verszeile-Tag (`<l>`) verwendet.
   - Markierung bei dem Brief – Ergänzung entsprechender Tags für die Grußformel (`<opener><salute>…</salute></opener>`) und für die Abschiedsformel (`<closer><salute>…</salute></closer>`).

10) Das Ergebnis wird zum Schluss des Hauptskriptes in die Zieldatei geschrieben (PTX: 274–276). Es handelt sich um eine XML-Datei, die im nächsten Schritt mit der XSLT-Schablone bearbeitet wird.

Die Datei ist jetzt eine wohlgeformte XML-Datei.[61] Die Markierung reicht bis zur Satzebene. Zur Rekapitulation der bisher verwendeten Tags (für den deutschen

60 Das Verb ‚auskommentieren‘ bedeutet ein Stück des Skripts als Kommentar zu markieren, damit dieser Abschnitt vom Computer nicht bearbeitet wird.

61 Die erste Seite des Romans mit dieser groben XML-Struktur befindet sich im Anhang Nr. 7.

Kratochvil-Text): nach den Regeln der XML-Struktur sind die Roman-Bücher (…), Kapitel (<chapter type="chapter">…</chapter>), Kapitelüberschriften (<head>…</head>), Absätze (…), Segmente (<seg>…</seg>), Sätze (<s>…</s>), Zeilen- (<lb/>) und Seitenumbruch (<pb/>) markiert. Die Gedichte stehen in den <lg>-Tags und die einzelnen Verszeilen in den <l>-Tags. Entsprechende Tags wurden auch bei dem Brief, d. h. den die Grußformeln am Anfang (<opener>…</opener>) und am Ende (<closer>…</closer>) des Briefes (beide jeweils in eigenem <salute>-Element) getaggt. Der ganze Brief steht in Tags <ersatz>…</ersatz>. Enthalten ist auch die Kodierung für die Formatierung der Kursivschrift (<hi rend="italic">…</hi>) und den Text in Kapitälchenschrift (<hi rend="smallCaps">…</hi>).

Neben den standardisierten TEI-Elementen werden hier also zwei Hilfsmarkierungen bzw. Ersatztags verwendet:

- <ersatz>…</ersatz> – dienen zur Bezeichnung des Briefes und
- <chapter>…</chapter> – markieren die Kapitel im Text

Diese Tags werden später in -Elemente umbenannt, die Verwendung der Ersatztags wurde deswegen eingeführt, damit sie bei der Bearbeitung mit der XSLT-Schablone nicht mit anderen -Elementen mitgezählt werden.

ad b) durch die XSLT-Schablone wird die Zählung den <lb/>-Tags und <pb/>-Tags hinzugefügt.[62] Die Tags , , <seg>, <s>, <lg>, <l> und <salute> bekommen eine eindeutige Identifizierung xml:id, die die Angaben zur Adressierung im Text und eine Autoren-ID enthalten. Ein Satz-Element sieht z. B. so aus: <s xml:id="div38.p496.seg1589.s1674_jku00de">, dies ist folgendermaßen zu lesen: „Es handelt sich um ein Satz-Element (<s . . .>), das sich im 38. Kapitel (div38), im Absatz Nr. 496 (p496) und im Segment Nr. 1589 (seg1589) befindet. Vom Anfang des Textes ist dieses der 1674. Satz (s1674). Es handelt sich um den Text ‚Jiří Kratochvil: Unsterbliche Geschichte' (jku) aus dem Jahr 2000 (00) und die Sprache ist Deutsch (de)." Ein Zeilenumbruch-Element sieht z. B. so aus: <lb n="187:7"/> und es bedeutet, dass es sich um die Seite Nr. 187 (die Zahl vor dem Doppelpunkt) und die siebte Zeile auf dieser Seite (die Zahl nach dem Doppelpunkt) handelt. Die

62 Die für den Kratochvil-Text verwendete XSLT-Schablone befindet sich im Anhang Nr. 8.

Nummerierung der Seiten und Zeilen nach der gedruckten Vorlage erlaubt es, die Referenzen zwischen dem Originaltext und dem elektronischen Korpus herzustellen (vgl. DUBININ / VADAYEV / SMOLSKAJA 2005, 265).

Auch die Ersatztags werden durch die XSLT-Schablone entsprechend behandelt: die Tags `<ersatz>…</ersatz>`, die zur Bezeichnung des Briefes dienen, werden in die Tags `<div>…</div>` ohne Attribut umgewandelt. Das `<chapter>`-Element, das jeweils die Kapitel im Text markiert, wird in das `<div>`-Element umgewandelt, gezählt und entsprechend als Kapitel attribuiert. Die Attribute sind im Unterkapitel über die `<div>`-Tags näher beschrieben (vgl. Unterkapitel „3.2.1 `<div>`“).

Die XSLT-Transformation wird mit dem Programm ‚<oXygen/> XML Editor‘ durchgeführt. Am Anfang der Schablone (`KR_nummerieren.xsl`) werden in zwei Variablen spezifische Werte angegeben, die den jeweiligen Text (den deutschen oder den tschechischen) betreffen. In der ersten Variablen wird mit der Autoren-ID bestimmt, um welches Buch es sich handelt. In der zweiten Variablen befindet sich der Startwert für den Seitenzähler, da er sich bei jedem Buch unterscheidet; wenn z. B. die erste Textseite Nr. 7 ist, startet der Zähler mit 6.

Im sog. Transformations-Szenario[63] wird festgelegt, welche Datei bearbeitet werden soll (`zieldatei_KR_de_format_xml.xml`) und wie die Ausgabe ist (der Datei-Name `zieldatei_KR_de_nummeriert.xml` und der Pfad, der in der Verzeichnisstruktur den Ort angibt, wo die Datei gespeichert werden soll) – in diesem Fall wird die Ausgabe ebenso als eine XML-Datei gespeichert.[64]

Die Zieldatei hat die Struktur einer wohlgeformten XML-Datei, es ist noch der TEI-Header hinzufügen, und das Ergebnis wird gegen ein Schema validiert.

ad c) Der Header wird in einer getrennten Datei vorbereitet, damit bei der Zählung mit der XSLT-Schablone keine Elemente aus dem Header mit den Elementen aus dem Text zusammengezählt werden (z. B. `<p>`-Elemente).[65]

Am Anfang der Gesamtdatei werden noch zwei Zeilen eingefügt:

---

63 ‚Szenario‘ ist Bezeichnung einer Funktion innerhalb des Programms Oxygen, das hier zur Bearbeitung der XML-Dateien verwendet wird.

64 Zur Illustration wird die erste Seite des Romans mit der finalen XML-Struktur im Anhang Nr. 9 angeführt.

65 Der Inhalt der Datei, die den Header für den deutschen Kratochvil-Text enthält (`header_krat_de_.xml`) befindet sich im Anhang Nr. 10.

`<?xml version="1.0" encoding="UTF-8"?>` – die Dokumentdeklaration, die die Version von XML und den Zeichensatz UTF-8 angibt.

`<?oxygen RNGSchema="myTEI.rnc" type="compact"?>` – mit dieser Zeile erstellt das Programm ‚<oXygen/>' die Verbindung zu dem Schema.

Die vollständige Datei wird abschließend gegen das für dieses Projekt erstellte Schema validiert, womit geprüft wird, ob die XML-Struktur den in dem Schema enthaltenen Regeln entspricht. Die valide Textdatei ist für das Verwenden im DeuCze-Korpus bereit.

3.4 Zusammenfassung

In diesem Kapitel wurde die Bearbeitung der durch das OCR-Verfahren erstellten Text-Datei beschrieben. Das Ziel ist es, den Text der Bücher in eine XML-Datei umzuwandeln. Die Auszeichnungssprache XML wurde im ersten Unterkapitel kurz vorgestellt. Das Prinzip der Auszeichnung besteht darin, dass bestimmte Textabschnitte mit Marken, mit den sog. Tags, versehen werden, um sie so mit entsprechenden zusätzlichen Informationen zu markieren.

Damit das XML-Dokument den Standards entspricht, muss die Datei korrekt zusammengestellt (wohlgeformt) sein, und die Struktur muss den Regeln der Kombinierbarkeit der Elemente und Attribute, die in einem Schema definiert werden, entsprechen (die Datei ist dann valide). Die einzelnen Elemente und Attribute und ihre Verwendung in diesem Projekt wurden im zweiten Unterkapitel beschrieben.

Im abschießenden Unterkapitel wurden die konkreten Schritte bei der Transformation der Texte in XML-Format am Beispiel der Bearbeitung des Romans von Kratochvil beschrieben. Die erstellten XML-Dokumente entsprechen den TEI P5 Empfehlungen für die Kodierung der Korpustexte.

Die Texte der behandelten Bücher (von Thomas Brussig und Jiří Kratochvil) sind bis zur Satzebene synoptisiert und über die Benutzeroberfläche des DeuCze-Korpus unter der Internet-Adresse ‚deucze.org' zugänglich.

Die in dieser Projektphase angewendeten Tags enthalten Angaben, die dem Korpusbenutzer auf unterschiedliche Weise vermittelt werden. Abgesehen von der Funktion der Tags, der Segmentierung und Formatierung des Textes zu dienen, sind

Adressierung im Werk und statistische Angaben vertreten. Auf der Benutzeroberfläche werden am Anfang des jeweiligen Absatzes Zahlen angeführt, die die Seiten- und Zeilennummern repräsentieren. Die Zahlen geben an, wo der konkrete Absatz jeweils im deutschen und im tschechischen Text beginnt. Diese Informationen werden beim Kopieren der Textausschnitte automatisch mitgeliefert, um die Textstellen genau zitieren zu können. Außerdem wird neben dem Text auch die Anzahl der Sätze im konkreten Absatz angezeigt, die der Kontrolle der Synoptisierung dient. Das Anzeigen der statistischen Angaben (wie z. B. Nummerierung der Kapitel, Absätze und Sätze) sind für die Oberfläche ausgeschaltet, können aber bei Bedarf aus dem Quelltext gewonnen werden.

Die Angaben, dass in einem Segment mehrere Sätze vorkommen, damit die Segmente im Verhältnis 1:1 angezeigt werden können, sind in der Standardeinstellung in der XML-Datei nicht vertreten. Sie können jedoch beim Interesse bei der Bearbeitung mit dem PHP-Skript entsprechenden `<seg>`-Elementen als Attribut eingeführt werden. Daneben enthalten viele Tags das Attribut `xml:id` mit der genauen Adressierung im konkreten Text. Um diese beiden Informationen für den Endbenutzer sichtbar zu machen, müsste auch das Synoptisierungsskript entsprechend angepasst werden.

In dieser Phase des Projektes war es das Ziel, nicht nur die Texte paarweise nach Sätzen abbilden zu können, sondern es wurde auch Wert auf die Möglichkeit gelegt, die untersuchten Textstellen präzise zitieren zu können. In Zukunft können die Texte zusätzlich mit Markierungen der Wortarten angereichert werden. Auf diese Weise kann die heute zugängliche Suche nach Zeichenketten um die Suche nach bestimmten grammatischen Konstruktionen erweitert werden. Für die folgende textlinguistische Untersuchung ist keine Sondermarkierung (wie z. B. Markierung der Referenzen) vertreten.

4 Textanalyse

In diesem Abschnitt werden die ersten Schritte der Analyse des ausgewählten Textes durchgeführt. Die Grundlage bildet der Text des Romans ‚Unkenrufe' von Günter Grass, es ist nämlich vorgesehen, die Analyse an einem Text durchzuführen, der Deutsch als Ausgangssprache hat. Dabei wird ein anderes Buch aus dem DeuCze-Korpus gewählt, das nicht zu den im ersten Teil dieser Arbeit vorgestellten Werken gehört, damit der Umfang dieses Korpus auch durch andere Texte repräsentiert wird. Die Untersuchung könnte selbstverständlich an jedem Text im DeuCze durchgeführt werden.

Ein Text wird hier als eine sprachliche Einheit verstanden, die aus Einheiten niedrigerer Ordnung, die durch unterschiedliche Mittel miteinander verbunden sind, besteht. Der Gegenstand eines Textes äußert sich im Textthema als der größtmöglichen Kurzfassung des Textinhalts. In der Regel wird das Textthema von Teilthemen begleitet (vgl. Brinker 2005, 55ff). Die Teilthemen spezifizieren das Textthema und sind ihm untergeordnet.

In einer späteren Phase der Analyse wird auf den Gegenstand des Textthemas eingegangen, und ein Teilthema wird im Textverlauf beobachtet. Es wird zuerst versucht festzustellen, welche Phänomene helfen können, das Teilthema im Text zu identifizieren. Die Wörter im Text bilden Beziehungen zueinander, die durch die Mittel der Kohäsion und Kohärenz beschrieben werden können. Diese Mittel können auch zeigen, wie das Teilthema im Text, das einen Referenzbereich bildet, im Laufe der Handlung eingeführt, wiederaufgenommen, weitergeführt und abgeschlossen und u. U. auch wiedereingeführt werden kann (vgl. Hausendorf / Kesselheim 2008, 105ff).

Der gewählte Text stellt eine abgeschlossene Einheit zur Untersuchung der sprachlichen Mittel dar, die einerseits die einzelnen Sätze zu dem eigentlichen Text verbinden und sich gleichzeitig an der Thema- bzw. Teilthemaentwicklung beteiligen.

Das Arbeitsverfahren der Textuntersuchung besteht aus mehreren Schritten, die größtenteils durch den Computer automatisch durchgeführt werden. Die Anordnung der folgenden Unterkapitel entspricht der Reihenfolge der bei der Untersuchung durchgeführten Schritte. In der ersten Phase wird versucht, mit den Mitteln der elektronischen Textverarbeitung das zu analysierende Teilthema im Text auszuwählen. Das Ausgangs-

kriterium für die Wahl des zu untersuchenden Teilthemas ist die Vorkommensfrequenz im Gesamttext. Die Frequenzangaben werden aus der erstellten Frequenz-Wortformenliste gewonnen. Die Wörter stehen im Text nicht isoliert, deswegen werden ausgewählte Arten der Wortschatzgruppierung aufgeführt, um die zwischenwörtlichen Beziehungen aufzugreifen. Der zweite Schritt der Untersuchung beschäftigt sich damit, wie die Beziehungen zwischen den Wörtern innerhalb eines Textes zu seiner Konstitution beitragen. Von dieser Erkenntnis geht dann die Analyse der Teilthemaentwicklung im Text aus.

4.1 Einleitende Schritte der Analyse

Als Ausgangspunkt für die Analyse dient eine Frequenz-Wortformenliste. Diese Liste zeigt jedoch keine Beziehungen zwischen den Wörtern an. Die Zugehörigkeit der Wörter im Text zueinander wird aufgrund verschiedener Gruppierungsarten des Wortschatzes festgestellt. Dabei werden auch die Verbindungsarten zwischen den Sätzen, also die Mittel der Kohäsion und Kohärenz, betrachtet.

4.1.1 Wortformenliste

Die Liste der im Grass-Text enthaltenen Wortformen, die nach ihrer Vorkommensfrequenz sortiert sind, bildet die Basis dafür, dass das zu untersuchende Teilthema im Gesamttext identifiziert wird. Diese Liste wird automatisch durch das Programm AntConc[66] aus dem untersuchten Text erstellt, der in einer Textdatei ohne weitere Formatierung gespeichert ist. Von den aufgelisteten Wortformen bleiben die folgenden Kategorien unbeachtet: Funktionswörter (Artikel, Präpositionen, Konjunktionen, Hilfsverben, usw.), Namen der Protagonisten (*Reschke*, *Alexandra*, *Brakup*, …) und andere nicht motivierte Wortformen bzw. auch Verwandtschaftsbeziehungen (*Paar*, *Ende*, *Witwe*, …). Die Doppelformen werden für die Untersuchungszwecke vollständig ausgeschrieben, dies betrifft die Einträge:

ab- und mitgehört
Abgas- und Schwefelgeruch
Alt- und Rechtstadt
Blumen- und Kranzdiebstähle
Blumen- und Kranzverkäufern
Einzel- und Doppelgräbern
Einzel- und Doppelzimmer
Einzel- und Massengräber
Erd- und Wechselkröten
für- und vorsorgend
Geburtstags- und Jubiläumsinserate
Grab- und Urnenfelder

66 Autor: Anthony Laurence, hier in der Version: 3.2.0u Linux – http://www.antlab.sci.waseda.ac.jp/antconc_index.html, zit.: 6. 2. 2010.

hin- und hergerissen
Kaiser- und Schichauzeiten
Kranken- und Sterbekassen (2 Mal)
Kur- und Fischerdorf
links- oder rechtshändig
Mit- und Ineinander
Mohn- und Kornblumen
Neben- und Unterhaltskosten
Neunzig- bis annähernd Hundertjährige
Orts- und Zeitangaben
Pacht- und Nutzungsverträge (2 Mal)
Park- beziehungsweise Friedhofseingang
Post- und Polizeibeamte
Schichau-, dann Leninwerft
Schrift- und Ornamentvergoldung
Überführungs- und Nutzungskosten
Weihnachts- und Neujahrsbrief
west- und süddeutschen
west- und südeuropäischen
Zeilen- noch Wortabstand
zwei- bis dreitausend

Die Häufigkeit der Wiederaufnahmen der vertretenen Belege zeigt die zentralen Textgegenstände (vgl. Brinker 2005, 57ff). Die einzelnen Einträge in der Liste verweisen auf die Isotopieknoten, die Isotopiebeziehungen können durch die Beobachtung des Inhalts und Kontextes entdeckt und rekonstruiert werden. Diese Angaben werden jeweils im Bezug auf den gesamten Textinhalt betrachtet.

Nach dem ersten, subjektiven Beobachten der Gesamtliste ist der erste relevante Beleg das Wort *Friedhofsgesellschaft*, das in dieser Form im Text 63 Mal vorkommt (ggf. noch zwei Mal in der Pluralform). Die Frequenz-Wortformenliste wird so sortiert, dass nur die Wortformen beibehalten werden, die zu dem Begriff *Friedhofsgesellschaft* oder zum Buchtitel in einem Zusammenhang stehen. Die sortierte Liste ist von der Kenntnis des Inhalts beeinflusst, deswegen steht ‚Friedhofsgesellschaft' nicht mehr an der ersten Stelle.

Die unten angeführten Einträge zeigen die ersten fünfzig Posten der sortierten Wortformenliste:

Rank	**Freq**	**Word**	357	19	Gelände	515	13	Bericht
101	72	Idee	364	19	Vergolderin	530	13	Hinweis
115	63	Friedhofs gesellschaft	388	18	Unke	558	12	Hektar
145	49	Friedhof	399	17	Erde	565	12	Konsistorial rat
216	33	Aufsicht	410	17	Priester			
224	32	Versöhnungs friedhof	454	15	Gesellschafter	572	12	Politik
227	31	Aufsichtsrat	455	15	Heimat	578	12	Sprache
258	28	Toten	470	15	Rücktritt	579	12	Steinpilze
285	24	Friedhöfe	475	15	Umbettung	580	12	Tote
290	24	Kirche	490	14	Friedhofs	581	12	Tätigkeit
315	23	Tod	504	14	Rede	598	11	berichtet
			506	14	Sitzung	607	11	Feld
			510	14	Verlauf	610	11	Friedhofs gelände
			511	14	Vizedirektor			

613	11	Grabsteine			ordnung	706	10	Sarg
624	11	Rathaus	673	10	geschäfts	722	9	Allerseelen
646	10	Aktion			führenden	728	9	Beerdigung
654	10	Angebot	675	10	Grab	731	9	Betrieb
656	10	Beerdigungen	686	10	Kröten			
670	10	Friedhofs	703	10	Personen			

Die Liste besteht aus drei Spalten: die Zahlen unter ‚Rank' geben die Position in der vollständigen Wortformenliste an, die Angaben bei ‚Freq' geben die Vorkommensfrequenz im Gesamttext an, und unter ‚Word' stehen die jeweiligen Wortformen.

Bei der Sortierung der Liste zeigt es sich, dass die im Text stark vertretenen Referenzbereiche diese sind: a) ‚Friedhof', b) ‚Gesellschaft' (im Sinne einer ‚Organisation') und c) ‚Idee' (als der Gedanke, die Friedhofsgesellschaft zu gründen) – alle drei können anhand des Kontextes im Referenzbereich bzw. Teilthema ‚Friedhofsgesellschaft' zusammengefasst werden. Konkrete Beziehungen sind im Text zu suchen, und erst der Kontext zeigt, dass z. B. zwischen ‚Friedhof' und ‚Parkgelände' eine Referenzbeziehung entsteht: „*Die Frage, ob das Versöhnungsfriedhof heißende Parkgelände zwischen der Technischen Hochschule und der Poliklinik eingezäunt werden sollte,* […]." (GGU, S. 152)[67] Der Kontext zeigt Beziehungen zwischen den Wörtern, die an den jeweiligen Text gebunden sind, es gibt aber auch Beziehungen zwischen den Wörtern, die von dem Text nicht abhängig sind.

4.1.2 Wortschatzgruppierung

Die Beschreibung der Wortschatzgruppierung ergänzt die kontextlose Wortformenliste, denn besonders die inhaltlichen Beziehungen zwischen den Wörtern machen die Referenzen zwischen den Textwörtern sichtbar. Die Kriterien der Wortschatzaufteilung sind formorientiert (Wortfamilien), inhaltsorientiert (onomasiologisches Paradigma) oder eine Kombination aus Form- und Inhaltsorientiertheit (Wortfeld) (vgl. Lutzeier 1995, 99ff).

Aus der Sicht der Formebene zeigt sich die Zugehörigkeit der Wörter zueinander in den Wortfamilien. „*Eine Wortfamilie ist eine Klasse von Wörtern, die allesamt mittels morphologischer Prinzipien aus einem Grundelement herleitbar sind.*" (Lutzeier

67 Die Abkürzung ‚GGU' referiert auf das bearbeitete Buch: Grass, Günter: *Unkenrufe*. 1992. Die Seitenangabe gibt die Seitennummer an, an der sich der zitierte Satz oder Textabschnitt befindet oder an der er beginnt.

1995, 99) Dieses Grundelement oder Basis ist in diesem Zusammenhang als Wurzelmorphem zu verstehen – die etymologische Verwandtschaft der Wörter verknüpft sie zu den Wortfamilien (vgl. SCHIPPAN 1992, 43).

Die flektierten Formen werden hier nicht einbezogen, entscheidender ist das Kriterium, dass unterschiedliche Wortarten vertreten sein können (vgl. LUTZEIER 1995, 100). Als Beispiel diene hier die Wortfamilie von „graben“ und die im Text enthaltenen Belege:

Ableitungen: *Grab*, *Grube*, *Gruft*, *begraben*, *Begräbnis*

Es werden auch Kompositionen aus Elementen der Wortfamilie mit anderen Elementen mitgezählt:

- Linkserweiterung:

  Doppelgrab, *Einzelgräber*, *Elterngrab*, *Engelsgrube*, *Erstbegräbnisse*, *Familiengrab*, *Fundgruben*, *Fürstengruft*, *Gemeinschaftsgräber*, *Goldgrube*, *Kindergräber*, *Klawittergrab*, *Massengräber*, *Pharaonengräbern*, *Sammelgräber, Tatarengräber, Totengräber*, *Urnengräber*, *Wassergräben*

- Rechtserweiterung:

  Begräbnisbetrieb, *Begräbnisbräuche*, *Begräbnisinstitut*, *Begräbniskosten*, *Begräbnispraxis*, *Begräbnisriten*, *Begräbnisrituale*, *Begräbniswesen*, *Begräbnisse, Grabbodenplatten*, *Grabensystems*, *Gräberfeld*, *Gräberreihe*, *Grabfeld*, *Grabhügeln*, *Grabkreuze*, *Grablegung*, *Grabmal*, *Grabmaleinfassung*, *Grabmalgeviert*, *Grabplatte*, *Grabrede*, *Grabreihe*, *Grabspruch*, *Grabstätte*, *Grabstein*, *Grabstellenreihe*, *Grabsteininschrift*, *Grabsteinfragmente*, *Grabsteingestaltung*, *Grabsteinplatte*, *Grabstelle*, *Gruftgewölbe*, *Grufthöhe*

- Links- und Rechtserweiterung:

  Bodengrabplatte, *Doppelgrabstelle*

Die Elemente innerhalb einer Wortfamilie sind etymologisch verwandt, doch nicht immer deutet die Ausdrucksseite auf die Verwandtschaft hin. Ob z. B. die Wörter *Frieden* und *Friedhof* zu einer Wortfamilie gehören, kann nur mithilfe eines etymologischen Wörterbuchs festgestellt werden.

Das Wort „*Friedhof*“ ist eine Zusammensetzung, die ursprüngliche Bedeutung ist „*eingehegter Raum*“[68], heute hat es die Bedeutung „*Begräbnisstätte*“. Das Bestimmungswort „*Fried-*“ gehört zum althochdeutschen „*vrīten*“ – „hegen“. Dagegen entstammt das Wort „*Frieden*“ dem althochdeutschen „*fridu*“ und bedeutet ursprünglich

68 Mehr zu den etymologischen Angaben: Duden. Das Herkunftswörterbuch. 2001, 237.

„*Schonung*, *Freundschaft*“. Deswegen gehören die Wortbildungen, die von „Frieden“ abgeleitet sind (*zufrieden*, *zufriedengeben*, *zufriedenstellend*, *friedfertig*, *friedlich*) zu einer eigenen Wortfamilie, im Unterschied zu Wörtern wie: *Friedhofsanlage*, *Friedhofsordnung*, *Urnenfriedhof, Dorffriedhof*, die zu ‚Friedhof‘ gehören.

Innerhalb der ‚onomasiologischen Paradigmen‘ können die Beziehungen zwischen den Wörtern auf der Inhaltsebene beobachtet werden (vgl. Lutzeier 1995, 101ff). Es handelt sich um Klassen von lexikalischen Elementen, die zu einem vorgegebenen Begriff gehören und einen Referenzbereich bilden. Dieser Begriff charakterisiert das ganze Paradigma (auch als Begriff- oder Sachgruppe bezeichnet), und die enthaltenen Elemente beteiligen sich inhaltlich an ihm. Die Zuordnung der Elemente in die jeweilige Gruppe hängt von der Auslegung der Zuordnungsfrage des Autors ab. Aus diesem Grund kann das als Ausgangspunkt gewählte Begriffssystem nicht als ein absolut gültiges Raster angesehen werden, obwohl diese Aufteilung durchaus behilflich ist.

Der letzte hier beschriebene Typ der Wortschatzgruppierung ist das Wortfeld. Die Wortfelder bezeichnen Gruppen sinnverwandter Wörter (vgl. Römer / Matzke 2005, 56ff). Zur Ermittlung der Wortfelder spielt die Intuition der Sprecher eine wichtige Rolle, doch es müssen genaue Kriterien aufgestellt werden. Bei den Untersuchungen früherer Sprachstufen (oder Untersuchungen einer Sprache als Nichtmuttersprachler) stellt die eigene Intuition keine verlässliche Quelle der benötigten Sprachdaten dar, und der Gebrauch in den Texten ist die einzige brauchbare Untersuchungsbasis. Für die weitere Arbeit ist die Erkenntnis wichtig, dass das Wortfeld zwischen dem Einzelwort und dem Gesamtwortschatz steht (vgl. Lutzeier 1995, 103ff).

Trier erwähnt in seiner Abhandlung zum Sinnbezirk des Verstandes (Trier, 1973), dass kein ausgesprochenes Wort im Bewusstsein des Sprechers und Hörers vereinzelt da steht, sondern immer eine Menge begrifflich enger oder ferner benachbarter Wörter auftaucht. Diese Begriffsverwandten bilden untereinander und mit dem ausgesprochenen Wort ein gegliedertes Ganzes, nämlich das Wortfeld, auch sprachliches Zeichenfeld genannt. Dieser inhaltlich zusammengehörige Teilausschnitt des Wortschatzes ist die äußere, zeichenhafte Seite der begrifflichen Aufteilung. Es wird die gegenseitige Abhängigkeit der Wörter im Feld betont und dass das Einzelwort seine inhaltliche begriffliche Bestimmtheit im Gefüge des Ganzen erhält.

Das Wortfeld ist ein Gebilde, das eine Formebene und eine Inhaltsebene aufweist. Die Formebene betrifft die Vorstellung der Wortfelder als eine spezielle paradigmatische Gruppierung. In den Vordergrund tritt die Möglichkeit der Substitution an einer ausgezeichneten Stelle im verbalen Kontext, deshalb wird als Kriterium für ein Wortfeld angeführt, dass die Elemente einer Wortart angehören (LUTZEIER 1995, 106ff).

> *Wenn er von den achtunddreißig bei Curicke aufgeführten Epitaphen in Sankt Marien **erzählte**, **gab** sie **Bericht** von ihrer vergoldenden Arbeit an einem lange verschollenen Epitaph aus dem Jahr 1588. **Sprach** er vom niederländischen Manierismus, **zählte** sie das halbe Pferd im roten Feld und die drei Lilien auf blauem Grund im Wappen des Jakobus Schadius als vergoldenswert **auf**. Er **lobte** die Anatomie der im Relief aus Gräbern auferstandenen Knochengerüste, sie **erinnerte** ihn an den goldenen Anfangsbuchstaben auf schwarzem Grund im unteren Breitoval.* (GGU, S. 68)

Die Inhaltsebene des Wortfeldes betrifft die semantische Ähnlichkeit der Elemente. Ein Beispiel für das Wörter-Paradigma ist: „*alt, plagiat, bestseller, interessant, neuheit, umfangreich, originell, dick, langweilig, teuer, erschienen, erhältlich, taschenbuch, verkauft, vergriffen, langatmig, neu, ladenhüter*“ (LUTZEIER 1995, 111). Nach der Angabe einer inhaltlichen Klammer, des Aspekts, wird das Paradigma konkretisiert. Z. B. deutet der Aspekt ‚Erwerbsmöglichkeit‘ auf das Paradigma „*alt, teuer, erschienen, erhältlich, verkauft, vergriffen, neu*“ hin. Je nach Lesart können die Wörter auch anderen Paradigmen zugeordnet werden. Durch die Aspekte werden allgemeine Inhalte vorgegeben, und diesen Inhalten müssen die Elemente (ihre Bedeutungen oder Lesarten) des jeweiligen Paradigmas entsprechen. Der Aspekt stellt das semantische Kriterium für die Bestimmung des Wortfeldes dar, unterschiedliche Autoren bezeichnen das Phänomen als Sinnbezirk, oder Feldwert, der die Lexeme in einem Wortfeld vereint. Die Felder sind inhaltlich strukturiert in Teilmengen (nach den Aspekten) und nach Sinnrelationen, wie z. B. ‚Gegensatz‘ oder ‚Gleichheit‘ (vgl. LUTZEIER 1995, 111f).

Diese Wortschatzeinteilungsarten helfen die Beziehungen zwischen den Wörtern im Text zu entdecken. Bei der Suche nach den relevanten Beziehungen zwischen den Textwörtern kann auch das Prinzip der Synonymie nützlich sein. Die Synonyme sind formal unterschiedliche sprachliche Einheiten, die eine ähnliche oder gleiche Bedeutung haben, sie sind semantisch identisch (z. B. *Friedhof* und *Gottesacker*). Dagegen sind Wörter wie *Schnittblumen* und *Dahlien* keine Synonyme, sie stehen in einer

Hyperonym-Hyponym-Beziehung. Weil sie sich aber auf das gleiche Objekt beziehen, sind sie referenziell identisch (SCHIPPAN 1992, 206f).

Das Ziel ist nicht, den gesamten Romanwortschatz nach bestimmten Kriterien zu sortieren, sondern mit den Regeln der behandelten Wortschatzgruppierungsarten die Beziehungen innerhalb des Referenzenbereichs des untersuchten Teilthemas zu ermitteln. Diese Beziehungen zwischen den Wörtern üben einen Einfluss auf den Textaufbau aus.

4.2 Textkonstitution

Ein Text ist keine zufällige Aneinanderreihung von Sätzen. Die Sätze sind miteinander auf unterschiedliche Weise verbunden und bilden so eine übergeordnete Einheit – den Text. Zwischen den Sätzen besteht ein Zusammenhang, der sich durch die Bedeutung der enthaltenen Wörter und Wortgruppen realisiert. Dieser Zusammenhang stellt die Ausdrucksseite eines Textes dar. Daneben entsteht auch die textuelle Inhaltsseite durch die Rekurrenz bestimmter Wortbedeutungen. Diese Wiederholung schafft einen inhaltlich-thematischen Bereich oder eine inhaltlich-thematische Ebene, die ‚Isotopie' genannt wird (s. WOLF 2008, 58f).

Der Text ist ein sprachliches Zeichen, das sich aus Zeichen niedrigerer Ordnung (wie Sätze und Wörter) zusammensetzt. Die Zeichen niedrigerer Ordnung verknüpfen sich mithilfe der syntaktischen Regeln zu den Zeichen höherer Ordnung. Auf diese Weise bestehen die Ausdrucksseiten der Zeichen höherer Ordnung aus den Inhaltsseiten von Zeichen unterer Ordnung (vgl. WOLF 2008, 60f).

Die Zusammenhänge und Beziehungen zwischen den textkonstituierenden Zeichen realisieren sich als Mittel der Kohäsion und Kohärenz (Isotopie).

4.2.1 Kohäsion

Die Beziehungen, die auf der Ausdrucksseite der sprachlichen Einheiten zu betrachten sind, werden als Kohäsion beschrieben:

> Die Kohäsion „betrifft die Art, wie die Komponenten des OBERFLÄCHENTEXTES, d. h. die Worte, wie wir sie tatsächlich hören oder sehen, miteinander verbunden sind. Die Oberflächenkomponenten *hängen* durch grammatische Formen und Konventionen *von einander ab*, so daß also Kohäsion auf GRAMMATISCHEN ABHÄNGIGKEITEN beruht." (DE BEAUGRANDE / DRESSLER 1981, 3f)

Die Wiederaufnahmen der Ausdrücke bilden also die konkreten Relationen zwischen den Sätzen und verbinden sie zu Texten. Die Kohäsion kann in zwei Relationsarten betrachtet werden, es sind die Topikrelationen und Konnexrelationen:[69]

Kohäsion in Topikrelationen

Die Wiederaufnahme eines Ausdrucks aus einem Vorgängersatz in einem (nicht immer unmittelbar folgenden) Nachfolgersatz wird als ‚Topik' bezeichnet. Die betroffenen Ausdrücke stehen im Verhältnis Substituendum (das Ersetzte) – Substituens (das Ersetzende). Nach der semantischen Relation der Topikpartner unterscheiden sich die einzelnen Topiktypen (vgl. Wolf 2008, 63):

1. Repetition identischer Lexeme

> *Dem Belag sind deutlich die Umrisse plattgewalzter **Kröten** eingezeichnet. Es ist nicht ein und dieselbe **Kröte**. Vier **Kröten** wurden nicht nur einmal, ich bin sicher, mehrmals überfahren.* (GGU, S. 159)

Um ein Substantiv als Wiederaufnahme zu bezeichnen, muss dieses das Merkmal ‚definit' tragen, z. B. durch Verwendung des bestimmten Artikels oder z. T. auch von Possessivpronomen (vgl. Brinker 2005, 29):

Ein Beispiel mit bestimmten Artikel:

> *Vor den Auslagen **einer Bäuerin**, die in einem Korb gehäuft und auf Zeitungspapier gebreitet Pilze, zudem in drei Eimern Schnittblumen anbot, fanden Witwer und Witwe einander. **Die Bäuerin** hockte seitlich der Markthalle zwischen anderen Bäuerinnen und dem Ertrag ihrer Kleingärten: Sellerie, kindskopfgroße Wruken, Lauch und rote Bete.* (GGU, S. 7)

Ein Beispiel mit dem Possessivpronomen:

> *Jedenfalls war die Witwe schon zur Stelle, als **der Witwer** anstieß, stolperte, doch nicht zu Fall kam. […] **Sein** Tagebuch bestätigt Allerseelen und gibt die Schuhgröße preis.* (GGU, S. 7)

2. Repetition durch Pro-Zeichen (Pronomina)

> *Ich hörte aus dem Geplapper der **Frauen** heraus, daß es ihr Wunsch war, nach nur kurzer Stadtrundfahrt die Große Allee von Danzig nach Langfuhr zu befahren, hin und zurück, wie dazumal, während Schülerzeiten.* [...] ***Sie** hätten diese Strecke oft per Fahrrad genommen.* (GGU, S. 26)

3. Topik mit lexematischer Variation

3.1. mit Synonymen

69 Aufteilung nach Wolf 2008, 62ff.

*Nach meiner Kenntnis von **Kröten**, die, zugegeben, nicht groß ist, dürften es Erdkröten sein; doch Reschke hat auf die Rückseiten der Fotos geschrieben: »Das war eine **Unke**« – »Diese **Unke** ruft nicht mehr« – »Platte **Unke**« – und »Noch eine plattgewalzte **Unke**, kein gutes Zeichen!« Mag sein, daß er recht hat. Da Rotbauch- und Gelbbauchunken kleiner als **Erdkröten** sind und er das Längenmaß der platten Körper mit fünf, zweimal mit fünfeinhalb und mit sechs Zentimetern angibt, werden es wohl doch Froschlurche, also **Unken**, gewesen sein und keine **Erdkröten**, die sich bis zu fünfzehn Zentimeter Körpermaß auswachsen.* (GGU, S. 160)

3.2. mit lexikalischer Inklusion

*»Kein Wunder«, schreibt der Witwer, »daß die Stände neben der Dominikshalle so dürftig bestellt aussahen, schließlich sind an Allerseelen **Blumen** gefragt.* [...] *Obgleich die **Dahlien** und **Chrysanthemen** mehr hergaben, entschied sich die Witwe für **Astern**.* (GGU, S. 8)

3.3. mit artgleichen Elementen

*Da **Rotbauch-** und **Gelbbauchunken** kleiner als Erdkröten sind und er das Längenmaß der platten Körper mit fünf, zweimal mit fünfeinhalb und mit sechs Zentimetern angibt, werden es wohl doch Froschlurche, also Unken, gewesen sein und keine Erdkröten, die sich bis zu fünfzehn Zentimeter Körpermaß auswachsen. Weil im Werden plattgewalzt, waren es sicher **Tieflandunken**.* (GGU, S. 160)

3.4. mit kontrastierenden Elementen

*Dabei geht ein Säkulum zu Ende, das sich Vernichtungskriegen, Massenvertreibungen, dem ungezählten **Tod** verschrieben hatte. Doch nun, mit Beginn des neuen Zeitalters, wird wieder das **Leben** ...«* (GGU, S. 15)

3.5. mit Wortbildungselementen

*Offenbar **Freundinnen**, die beiden. Sie hätten diese Strecke oft per Fahrrad genommen. Übrigens sagten sie alle Gassennamen und Ortsbezeichnungen wie altgewohnt auf; und Mister Chatterjee verstand.« Der Bengale setzte sich eine Rennfahrerkappe auf. Reschke wünschte gute Fahrt. Die beiden **Schulfreundinnen** riefen: »Können wir Ihnen wärmstens empfehlen, dieses Vergnügen!* (GGU, S. 59)

Die Kohäsion mit Wortbildungselementen wird im Unterkapitel „4.2.3 Wortbildung und Text“ ausführlicher behandelt.

3.6. Topiks durch Paraphrasen

3.6.1. (Begriffs-)Expansion

*Wie nebensächlich steht das geschrieben, und doch wird ihm dieser **Computer**, den er, laut Selbstzeugnis, »nur stümperhaft« zu bedienen versteht, bei der Fleischwerdung ihrer Idee behilflich werden. **Ein sogenannter PC,***

wahrscheinlich Marke »Apple«. (GGU, 93)

3.6.2. (Begriffs-)Kondensation

Er sagte: ***»Was wir Heimat nennen, ist uns erlebbarer als die bloßen Begriffe Vaterland oder Nation, deshalb haben so viele, gewiß nicht alle, doch mit dem Älterwerden eine wachsende Zahl Menschen den Wunsch, sozusagen zu Haus unter die Erde zu kommen, ein Wunsch übrigens, der zumeist bitter unerfüllt bleibt, denn oft stehen die Umstände diesem Verlangen entgegen. Wir aber sollten von einem Naturrecht sprechen. Im Katalog der Menschenrechte müßteendlich auch dieser Anspruch verbrieft sein. Nein, nicht das von den Funktionären unserer Flüchtlingsverbände geforderte Recht auf Heimat meine ich – die uns eigentümliche Heimat ist schuldhaft und endgültig vertan worden –, aber das Recht der Toten auf Heimkehr könnte, sollte, dürfte angemahnt werden!«*** *Ich vermute, daß Professor Dr. Reschke diesen* ***Vortrag*** *sowie weitere Gedanken* […]. (GGU, S. 37)

4. Elliptische Topiks

Er winkte mich herüber, bis ich seine Rikscha, denn sie gehörte ihm, bewundern konnte. ***Alles blitzblank****.* (GGU, S. 57)

Die Topiks, die die Ausdrucksseite des Textes vertreten, formen die Isotopien des Textes, „*d. h. die Kontinuität und das Fortschreiten des Inhalts*“ (vgl. Agricola 1975, 28) und verknüpfen so die Sätze zu Texten.

Kohäsion in Konnexrelationen

Die im Text enthaltenen Funktionszeichen erzeugen Konnexrelationen zwischen den Sätzen. Die semantische Relation wird durch Konjunktionen, Konjunktional-adverbien und auch festen Zeichenfolgen (wie *kurz und gut*) signalisiert (s. Wolf 2008, 63f). In dieser Arbeit richtet sich das Hauptuntersuchungsinteresse auf die Kohäsion in Topikrelationen, deswegen wird die Kohäsion in Konnexrelationen nicht weiter behandelt.

4.2.2 Kohärenz

Die sprachlichen Einheiten (z. B. die Wörter) werden auf der Inhaltsseite durch die Mittel der Kohärenz zu größeren Einheiten verbunden:

> Die Kohärenz „betrifft die Funktionen, durch die die Komponenten der TEXTWELT, d.h. die Konstellation von KONZEPTEN (Begriffen) und RELATIONEN (Beziehungen), welche dem Oberflächentext zugrundeliegen, für einander *gegenseitig zugänglich* und *relevant* sind.“ (de Beaugrande / Dressler 1981, 5)

Ähnlich wie auf der Ausdrucksebene die Topiks und Topikketten gebildet werden, werden die auf diese Weise verbundenen Elemente auf der Inhaltsebene zu sog. Isotopien und Isotopieketten zusammengefasst.

> Isotopie: Eine auf A. J. Greimas (1966) zurückgehende Bez. für eine Form der Bedeutungsbeziehung zwischen den Lexemen eines Textes, die auf semant. Äquivalenz beruht und erklärt wird als wiederholtes Vorkommen von Semen (Semrekurrenz) in unterschiedl. lexikal. Einheiten des Textes, z. B. *Der Großvater... der alte Herr... er... seine erste Liebe... der Graukopf...* usw. (Metzler Lexikon Sprache 2000, 320)

Die Isotopien und Isotopieketten bzw. das Vorkommen mehrerer Isotopieketten im Text bilden das Isotopienetz und konstituieren so das Textthema.

Die Kohäsion macht sich auch durch die Mittel der thematischen Progression sichtbar. Das Fortschreiten vom ‚Thema' zum ‚Rhema' im Sinne der funktionalen Satzperspektive spielt sich auf dem Hintergrund von Kontext ab. Die Thema-Rhema-Struktur, als inhaltsseitiges Element, kooperiert mit den ausdrucksseitig fungierenden Topiks und ergibt die ‚thematische Progression'. Ein ‚Thema' ist für den Hörer der Anknüpfungspunkt an Bekanntes, schon Erwähntes. Über das ‚Thema' wird das ‚Rhema', das Neue, noch nicht Erwähnte, ausgesagt. (Wolf 2008, 64f)

> *Vor den Auslagen **einer Bäuerin**, die in einem Korb gehäuft und auf Zeitungspapier gebreitet Pilze, zudem in drei Eimern Schnittblumen anbot, fanden Witwer und Witwe einander. **Die Bäuerin** hockte seitlich der Markthalle zwischen anderen Bäuerinnen und dem Ertrag ihrer Kleingärten: Sellerie, kindskopfgroße Wruken, Lauch und rote Bete.* (GGU, S. 7)

Im ersten Satz ist das Thema ‚die Auslagen', vor denen sich das Paar trifft, erst im zweiten Satz rückt ‚die Bäuerin' in den Vordergrund und wird selbst zum Thema.

Dagegen bei der semantischen Progression wird das semantische Verhältnis zwischen den Sätzen durch Konnektoren explizit ausgedrückt, doch die Konnektoren werden im Text nur selten verwendet. Durch die Abfolge der Sätze wird die semantische Relation zwischen einem Vorgänger- und dem Nachfolgersatz aufgrund des Kontextes deutlich, diese Beziehung kann als semantische Progression bezeichnet werden (Wolf 2008, 67f), z. B.:

Bedingung – Folge: *Er zahle gut. Das Ganze sei ausbaufähig.* (GGU, S. 58)

Beabsichtigte Handlung – Begründung: *Er sagte, seine Gesellschaft plane, den Mitgliedern der Versicherung längere und kürzere Kuraufenthalte in Osteuropa*

insbesondere in den ehemals deutschen Provinzen zu ermöglichen. An Nachfrage fehle es nicht. (GGU, S. 55)

Die semantischen Beziehungen können die Satzgrenze überschreiten, und dank der gemeinsamen Seme können die Elemente der Referenzbereiche im Textverlauf identifiziert werden, denn ihre Referenzidentität spielt bei weiterer Bearbeitung eine wichtige Rolle.

Der Zusammenhang auf der Ausdrucksseite des Textes wird durch die lexikalischen Wiederaufnahmen gebildet. Der inhaltsseitige Zusammenhang wird in den Isotopieketten, die den Text inhaltlich-thematisch bestimmen, realisiert. Dabei ist bei der weiteren Bearbeitung der Zusammenhang zu einem bestimmten Referenzbereich wichtig. Mit der Kohäsion im Text ist die Wortbildung eng verbunden.

4.2.3 Wortbildung und Text

Die Wörter als Wortbildungsprodukte stellen die Bausteine des Textes dar, gleichzeitig setzen die Texte die Rahmenbedingungen für die Bildung und den Gebrauch von Wörtern. Besonders wichtig sind hier die textkonstitutiven Funktionen Textverflechtung und thematische Entfaltung. Die Wortbildung hilft, die Text-Kohäsion zu bilden, diese Eigenschaft beruht auf der morphosemantischen Motiviertheit der Wortbildungsprodukte und zeigt sich in der Wiederaufnahme ein und desselben Grundmorphems in mehreren Wörtern. Die Wörter, die dasselbe Grundmorphem enthalten, können zu einer Isotopiekette gehören und auf diese Weise die Kohärenz eines Textes verdeutlichen (vgl. Barz [u. a.] ^{3}2004, 60).

Unter Verwendung der Frequenz-Wortformenliste (vgl. Unterkapitel „4.1.1 Wortformenliste“) konnte festgestellt werden, dass die Referenzen *Friedhofsgesellschaft* bzw. *Friedhof* im ganzen Text sehr oft vorkommt, sie bauen ein semantisches Netz über dem Roman auf. Das Ziel der Analyse ist es herauszufinden, mit welchen konkreten Mitteln das Teilthema, das den Referenzbereich bildet, wieder aufgenommen wird. In erster Linie werden diese Mittel aus der Sicht der Wortbildung betrachtet.

Das Wort *Friedhofsgesellschaft* ist ein Dekompositum (eine Art von Determinativkompositum, wo eine Konstituente, hier *Friedhof*, wiederum ein Kompositum ist), das sich aus dem Bestimmungswort *Friedhof* und dem Grundwort *Gesellschaft* (verbunden mit dem Fugenelement *s*) zusammensetzt. Durch die Assoziationen mit den

Bestandteilen können weitere Mittel der Teilthemaentwicklung entdeckt werden. Im Rahmen der Zugehörigkeit der Elemente zueinander müssen semantische Beziehungen (Wortfelder) oder Beziehungen aufgrund der Wortbildung (Wortfamilien) erwähnt werden. Zuerst wird die Wortbildungsaktivität der ersten Komponente, also *Friedhof*, betrachtet.

Weil die Wortbildungen, die von *Frieden* abgeleitet sind (*zufrieden*, *zufriedengeben*, *zufriedenstellend*, *friedfertig*, *friedlich*) zu einer selbstständigen Wortfamilie gehören, werden sie deswegen aus der folgenden Untersuchung ausgelassen.

Die Vorkommensformen von *Friedhof*[70] werden in einer Übersicht[71] nach der Wortbildungsart geteilt:

a) als einfache Wortbildung – Kompositum:

Friedhof

b) *Friedhof* als erste unmittelbare Komponente innerhalb eines Dekompositums, wobei jeweils nach der Art der zweiten unmittelbaren Komponente (UK) können die Dekomposita nach gemeinsamen Elementen gruppiert werden:

- die zweite UK umschreibt den Friedhof:

Friedhofsanlage; *Friedhofsgelände*; *Friedhofshügel*

- die zweite UK gibt das Personal an:

Friedhofsgärtner

- die zweite UK gibt das an, was mit dem Gelände des Friedhofs verbunden ist:

Friedhofsalleen; *Friedhofsbäume*; *Friedhofseingang*; *Friedhofsfelder*; *Friedhofslinden*; *Friedhofsportal*; *Friedhofsrondell*; *Friedhofszaun*

- die zweite UK gibt das Organisatorische des Friedhofs an:

Friedhofsangebot; *Friedhofsbetrieb*; *Friedhofsgesellschaft*; *Friedhofsordnung*; *Friedhofstätigkeit*; *Friedhofsverwaltung*

- die zweite UK beschreibt die Menschen, die den Friedhof zu einem bestimmten Zweck betreten:

Friedhofsbesuch; *Friedhofsbesucher*

70 Die vorhandenen flektierten Formen werden nur dann angegeben, wenn im Text keine Grundform steht.

71 Diese Übersicht liegt in ihrer grafischen Form im Anhang Nr. 11 bei.

- die zweite UK beschreibt die Atmosphäre auf dem Friedhof:

Friedhofsruhe; *Friedhofsstimmung*

- die zweite UK beschreibt die Tätigkeit, die auf dem Friedhof stattfindet:

Friedhofsgespräch

- die zweite UK enthält Ausdrücke, die die Räumlichkeiten des Friedhofs (wie Lokation und Anordnung) betreffen:

Friedhofslage; *Friedhofsplänen*

- die zweite UK betrifft den Gedanken, die Friedhofsgesellschaft zu gründen / zu betreiben:

Friedhofsidee

c) *Friedhof* als zweite unmittelbare Komponente (UK) innerhalb eines Dekompositums, wobei jeweils nach der Art der ersten UK können die Dekomposita nach gemeinsamen Elementen gruppiert werden::

- die erste UK ist von dem Pfarreinamen abgeleitet:

Barbarafriedhof; *Marienfriedhof*; *Salvatorfriedhof*; *Katharinenfriedhof*; *Mennonitenfriedhof*

- die erste UK gibt an, wo sich der Friedhof befindet:

Dorffriedhof, *Waldfriedhof*

- die erste UK gibt den Gründungszweck des Friedhofs an:

Versöhnungsfriedhof

- die erste UK gibt an, welcher Bevölkerungsgruppe der Friedhof dient:

Armenfriedhof; *Militärfriedhof*; *Soldatenfriedhof*; *Garnisonsfriedhof*

- die erste UK gibt die Art an, wie die sterblichen Überreste bestattet werden:

Urnenfriedhof

- die erste UK gibt die weitere Ausstattung des Friedhofs an:

Krematoriumsfriedhof

- die erste UK gibt weitere Sonderangaben an:

Extrafriedhof; *Spezialfriedhof*

Nach den Angaben in der Wortformenliste ist das Wort *Friedhof* statistisch folgendermaßen vertreten:

- als alleinstehend 95 Mal (alle Tokens[72]: *Friedhof*, *Friedhöfe*, *Friedhofs*, *Friedheefe*[73], *Friedhöfen*),

- als 1. UK eines Dekompositums 192 Mal (z. B. in den Formen: *Friedhofsgesellschaft*, *Friedhofsgelände*, *Friedhofsordnung*)

Beispiel für Topik mit Wortbildungselementen:

> *Sie wollte keine **Friedhöfe** mehr sehen. Ihre zu leichten Schuhe seien für weiteres **Friedhofsgelände** ungeeignet.* (GGU, S. 225)

- als 2. UK eines Dekompositums 64 Mal (z. B. in den Formen: *Versöhnungsfriedhof*, *Marienfriedhof*, *Waldfriedhof*)

Beispiel für Topik mit Wortbildungselementen:

> *Beifällig wurden die Hinweise der Piątkowska auf das Fernziel Wilna und den dort in Pacht zu nehmenden **Friedhof** bewertet. Der könne gleichfalls **Versöhnungsfriedhof** heißen, denn zwischen Litauen und Polen bestehe Bedarf an Versöhnung.* (GGU, S. 120)

- innerhalb einer Abkürzung 5 Mal (als *PDLFG* und *DPFG*)

Beispiel für die Wiederaufnahme:

> *Und diesen frühen Einübungen entsprechend, begann der Professor der Kunstgeschichte, mit dem Kapital der Deutsch-Polnischen **Friedhofsgesellschaft** zu arbeiten. Am Aufsichtsrat vorbei, hat er Gespür bewiesen, Konten gesplittet, hier rechtzeitig abgestoßen, dort Gewinn gemacht und unterm Strich Grundlagen für Investitionen geschaffen, mit deren Hilfe die **DPFG** ein verzweigtes, ich meine, ein undurchsichtiges Unternehmen wurde.* (GGU, S. 170)

Die bisher besprochene Auflistung behandelt die im Text vorkommenden Formen, ohne eigene Beziehungen zwischen ihnen. Die Wörter stehen im Text nicht isoliert, und die einzelnen Satzglieder enthalten oft mehrere Einheiten. Die Wortformenliste nimmt keine Rücksicht auf den Kontext, deswegen wird noch eine Untersuchung der Kontextumgebung von *Friedhof* durchgeführt. Diese Untersuchung soll zeigen, mit welchen Wörtern das Lexem *Friedhof* im Text am häufigsten kombiniert wird. Die Unterscheidung der sich deckenden Kasusformen spielt bei der

72 Unterscheidung der Einträge nach Gesamtanzahl (‚Token') und zusammengefasste Gruppierungen (‚Type'), vgl. McEnery/Wilson ^{2}2001, 82.

73 Das Wort *Friedheefe* wurde in dieser Form aus dem Text (GGU, S. 138) übernommen.

Sortierung keine Rolle, doch anhand der unterschiedlichen Form wird zwischen Singular und Plural unterschieden.

Friedhof in der Singularform kommt in unterschiedlichen Kasus 63 Mal vor. Davon sind 26 Vorkommen alleinstehend und der Rest sind Wortkombinationen, *Friedhof* ist also erweitert durch:

a) Ortsangabe mit Präposition (6 Mal): *Friedhof am Hagelsberg, Friedhof in Wilna, Friedhof in Wilno*

b) Adjektiv, das von der Staats-Zugehörigkeit abgeleitet ist (7 Mal): *deitsche Friedhof*[74]*, deutsche Friedhof, polnische Friedhof*

c) Kombination der beiden vorher genannten (3 Mal): *deutsche Friedhof in Gdańsk, Polnische Friedhof in Wilna, polnischen Friedhof in Wilno*

d) die Kombinationen mit unterschiedlichen Adjektiven (14 Mal) sind zwar häufig vertreten, aber die Adjektive wiederholen sich seltener, Beispiele: *groß, katholisch, anschließend, alt, zukünftig, stillgelegt, einstig*

e) Sonstiges (7 Mal), wie z. B. die Bezeichnung mit Eigennamen bzw. mit Pfarreinamen: *Friedhof Matarnia*, *Sankt-Joseph-Friedhofs*, oder die Angabe, für welche Bevölkerungsgruppe der Friedhof bestimmt ist: *Friedhof für Deutsche*

Friedhof in der Pluralform kommt in unterschiedlichen Kasus 32 Mal und davon 8 Mal als alleinstehend vor. Der Rest sind Wortkombinationen:

a) Ortsangabe mit Präposition (einmal): *Friedhöfen in Gdańsk*

b) Adjektiv, das von Staats-Zugehörigkeit abgeleitet ist (2 Mal): *deutsche Friedhöf*e, *norddeutschen Friedhöfen*

c) die Kombinationen (21 Mal) mit unterschiedlichen Adjektiven, Beispiele: *vereinigt, abgeräumt, erweitert, eingeebnet*

Aus der Gruppe der unterschiedlichen Adjektive kommt *vereinigt* 16 Mal vor, davon 15 Mal mit Großbuchstaben als Eigenname von *die Vereinigten Friedhöfe* (einmal als *Väainichte Friedheefe*[75]) und einmal mit Kleinbuchstaben als allgemeine Bezeichnung der Vereinigung mehrerer Friedhöfe.

74 Das Wort *deitsche* wurde in dieser Form aus dem Text (GGU, S. 130) übernommen.
75 Die Wörter *Väainichte Friedheefe* wurde in dieser Form aus dem Text (GGU, S. 138) übernommen.

Das Wort *Friedhof* kommt oft als eine Komponente eines Dekompositums vor. Insgesamt sind es 129 Tokens für *Friedhof* als erste unmittelbare Komponente (wie in *Friedhofsordnung*) und 64 Vorkommen für *Friedhof* als zweite unmittelbare Komponente (wie in *Versöhnungsfriedhof*).

Es folgen die jeweils am häufigsten vorkommenden Typen dieser im untersuchten Text vertretenen beiden Gruppen:

- *Friedhofsgesellschaft* kommt im Text 63 Mal vor, dabei 19 Mal *Deutsch-Polnische Friedhofsgesellschaft*, 1 Mal als *Polnisch-Deutsche Friedhofsgesellschaft* und 4 Mal *Polnisch-Deutsch-Litauische Friedhofsgesellschaft* und 39 Mal mit anderen Adjektiven oder an der zweiten Stelle innerhalb eines substantivischen Attributs (einer Genitivverbindung zweier Substantive).

- *Versöhnungsfriedhof* kommt im Singular 32 Mal vor, dabei gibt es keinen Beleg mit adjektivischen Attributen, sondern es steht alleine als Bezeichnung für den von den Protagonisten gegründeten Friedhof. Im Text gibt es 3 Belege für Pluralform (*Versöhnungsfriedhöfe*), bei denen es nur einmal zu einer Verbindung mit Attribut kommt: *die neuentstandenen Versöhnungsfriedhöfe*.

Es hat sich gezeigt, dass die Suche nach den Belegen nach den Regeln der Wortbildung mithilfe des Computers ohne großen Aufwand betrieben werden kann.

Im Hinblick auf diese Tatsache wurde versucht, die Beziehungen zwischen den Wörtern des untersuchten Textes zu visualisieren und so einen Überblick über die im Text vertretenen Belege zu bekommen. Im Anhang Nr. 12 findet sich ein Bild, das einen (manuell erstellten) Vorschlag einer derartigen Darstellung bietet. Es könnte nützlich sein, ein Programm zu bilden, welches diese Sortierung und Abbildung automatisch erstellen könnte. Dabei wären Informationen einzugeben wie: der Quelltext, der Ausgangsausdruck, seine Synonyme und andere auf ihn bezogene Ausdrücke. Der Wortbestand würde anhand seiner Ausdrucksseite untersucht und die betroffenen Treffer entsprechend sortiert. Die Ausgabe bringt eine Übersicht über den konkret vertretenen Wortschatz, eine Referenz auf die Textstelle und Kontextangaben. Diese Visualisierung könnte ein sinnvolles Instrument einer Textanalyse sein, das eine vorläufige Übersicht über das zu untersuchende Material gibt.

4.3 Zusammenfassung

In diesem Kapitel wurden die Instrumente und die Vorgehensweise beschrieben, die bei der Textanalyse als Hilfsmittel verwendet werden können.

Das erste Unterkapitel behandelte die Wortformenliste, die dazu dient, auf das zu untersuchende Teilthema des Textes mit konkreten Belegen hinzuweisen. Diese Belege wurden durch ihre Vorkommensfrequenz im Gesamttext sortiert. Die Liste gibt keine Angaben zu den Beziehungen zwischen den Wörtern im Text. Diese Beziehungen werden in dieser Phase anhand der ausgewählten Arten der Wortschatzgruppierung identifiziert. Beim Bestimmen der konkreten Referenzen ist der Kontext entscheidend. Im zweiten Unterkapitel wurde untersucht, wie sich die Wörter im Text aufeinander beziehen und wie sie sich als Mittel der Kohäsion und Kohärenz an der Textkonstitution beteiligen. Das wiederholte Vorkommen bestimmter Kohäsions- und Kohärenzmittel wird in dieser Arbeit dazu verwendet, die Textuntersuchung durch das automatische Auffinden der Belege durch den Computer entsprechend vorzuentlasten.

Unter Anwendung der in diesem Kapitel gewonnenen Erkenntnisse wird im Folgenden versucht, eine Untersuchung der Teilthemaentwicklung durchzuführen.

5 Textthema und Teilthemaentwicklung

Dieses Kapitel konzentriert sich auf die Analyse der Entwicklung eines Teilthemas im Gesamttext. Zuerst wird der Begriff ‚Textthema' erläutert, nachfolgend werden die Themahinweise im Text behandelt und abschließend wird die Analyse der Teilthemaentwicklung durchgeführt.

5.1 Textthema

Das Thema eines Textes besagt, wovon der Text handelt. Im alltäglichen Verständnis stellt das Thema den Gegenstand eines Textes und das, was über diesen zentralen Gegenstand ausgesagt wird, dar. Das Thema wird hier als Kern des Textinhalts verstanden, und es kann in einem Textsegment (wie in einer Überschrift) angegeben oder durch eine zusammenfassende Paraphrase des Textinhalts ermittelt werden. Ein Text enthält in der Regel mehrere Themen mit einer unterschiedlichen thematischen Relevanz, deswegen kann zwischen dem Hauptthema und Nebenthemen differenziert werden (vgl. Brinker 2005, 55f). Hackl-Rössler betrachtet das Thema als „eine semantische, nicht-sprachliche Größe"[76]. Die Themen und Teilthemen können durch Zusammenfassungen oder Schlüsselwörter, die die betreffenden Textabschnitte vertreten, versprachlicht werden. Die Zusammenfassungen sind an das Textverständnis des jeweiligen Lesers gebunden, deswegen ist die interpretative Themenbestimmung immer auch subjektiv, denn jeder Leser kann den Text (teilweise) anders verstehen (Vgl. Hackl-Rössler 2006, 52ff).

Nach Brinker[77] wird das Thema als Gesamtinhalt des Textes durch die Verknüpfung bzw. Kombination relationaler, logisch-semantisch definierter Kategorien entfaltet. Unter diesen Kategorien sind die Beziehungen zwischen den Teilinhalten bzw. Teilthemen, die sich in einzelnen Textteilen realisieren, zu verstehen. Die Teilthemen spezifizieren das Hauptthema, sind ihm untergeordnet und können von dem Hauptthema abgeleitet werden (vgl. Brinker 2001, 57).

Das Gesamttextthema befindet sich also an der Spitze der Thema-Hierarchie und kann durch untergeordnete Teil- bzw. Subthemen gegliedert und spezifiziert oder durch

76 Siehe Hackl-Rössler 2006, 52.
77 Vgl. Brinker 2005, 61ff.

gleichrangige Nebenthemen ergänzt werden (vgl. HACKL-RÖSSLER 2006, 55ff). Eine weitere Stufe der Hierarchie der Themeneinteilung bezeichnet BRINKER als Unterthemen.[78]

Eine andere Art der Beschreibung der inhaltlichen Beziehungen zwischen einzelnen Teilen des Textes stellt die Quaestio dar. Der Text beantwortet in seiner Gesamtheit eine Frage, die Quaestio des Textes. Dies betrifft auch einzelne Äußerungen. Nicht jeder Äußerung geht eine explizite Frage voraus, man kann sich eine implizite Frage hinzudenken. Die Antwort auf diese Frage kann sich auch auf eine Reihe von Äußerungen verteilen, die in bestimmter Weise miteinander verknüpft sind. Diese Verknüpfung wird durch Verwendung bestimmter Ausdrucksmittel deutlich, die den äußeren Ausdruck für die Art und Weise, wie die Informationen schrittweise eingeführt und entfaltet werden, darstellen. Jede Äußerung enthält ein Gefüge von Angaben zu Ort, Zeit, Handlung, Personen und Modalität zu verschiedenen semantischen Bereichen, den Referenzbereichen. Die Entfaltung der Informationen von Äußerung zu Äußerung wird als referentielle Bewegung bezeichnet (vgl. KLEIN, 1987, 163ff).

Die weitere Textuntersuchung geht von konkreten Belegen im Text aus, um das ausgewählte Teilthema und seine Progression zu beobachten. Eine vollständige Untersuchung der Bereiche anhand der Quaestio würde den Rahmen dieser Arbeit sprengen, deswegen wurde sie nicht durchgeführt.

Für die weitere Arbeit wird von der folgenden hierarchischen Anordnung der Themen ausgegangen:

- Textthema als Thema des Gesamttextes,
- Teilthema, das dem Textthema untergeordnet ist, und
- das Subthema als die Aufteilung und Spezifizierung der Teilthemen.

Auf das Thema wird mit bestimmten Elementen, die im Text implizit oder explizit enthalten sind, hingewiesen.

5.2 Themahinweise

Die Thematik eines Textes besteht in der thematischen Zusammengehörigkeit der sprachlichen Erscheinungsformen im jeweiligen Text. Im Text kommen also bestimmte thematische Hinweise vor, die dem Leser (oder Hörer) signalisieren, worum es im Mitgeteilten geht. Wenn sie dazu beitragen, die Elemente eines Textes miteinander

78 Vgl. BRINKER / HAGEMANN 2001, 1254ff.

zu verbinden, können sie auch in der Rolle der Verknüpfungshinweise (wie Rekurrenz oder Pronominalisierung) auftreten (s. Hausendorf / Kesselheim 2008, 27). Die verknüpfende Funktion der Hinweise bzw. die Kohäsionsbeziehungen helfen, die Elemente eines Teilthemas als Bestandteil des Themas im Text zu identifizieren.

Brinker weist darauf hin, dass „*die textanalytische Bestimmung des Themas primär auf interpretativen Verfahren beruht*“[79]. In dieser Arbeit wird bei der Identifizierung der Themenhinweise so vorgegangen, dass zuerst die Vorkommenshäufigkeit der jeweiligen Belege bewertet und durch anschließende subjektive Bewertung bestätigt oder widerlegt wird, ob der jeweilige Ausdruck als ein relevanter Themahinweis zu beurteilen ist. Als Beispiel sei hier die Ausscheidung der Funktionswörter oder Eigennamen von der eigentlichen Untersuchung genannt. Obwohl hier von maschineller Textverarbeitung ausgegangen wird, ist die Themabestimmung schließlich von dem Gesamtverständnis des Textes abhängig.

Als ein Themahinweis wird im Text alles betrachtet, was dazu beiträgt, die Zusammengehörigkeit der in einem Text auftretenden referenziellen Hinweise entstehen zu lassen (s. Hausendorf / Kesselheim 2008, 103).

Hausendorf und Kesselheim führen Kriterien an, die bei der Untersuchung der Themahinweise in Betracht gezogen werden müssen[80]:

1) Themahinweise hängen mit referenziellen Hinweisen zusammen, deswegen müssen die wort- und satzbezogenen Referenzen berücksichtigt werden.

2) Der thematische Gesamtzusammenhang eines Textes entwickelt sich mit dem Voranschreiten der Lektüre.

3) Die Bestimmung der Textthematik(en) sollte aus der Rekonstruktion der Themahinweise möglichst textnah hervorgehen. Das Ziel ist die Identifizierung der für den Aufbau der Textthematik relevanten Hinweise im Text, nicht ihre Paraphrase.

4) Texte können ihre Thematik metakommunikativ benennen und erläutern. Unter Umständen können die Texte aber ihren thematischen Zusammenhang gezielt im Unklaren lassen oder erlauben mehrere Lesarten.

79 Siehe Brinker 2005, 56.

80 Siehe Hausendorf / Kesselheim 2008, 103f. Die Beschreibung der Punkte wurde teilweise wörtlich übernommen.

5) Themahinweise fallen in vielen Fällen mit Verknüpfungshinweisen (z. B. Pronominalisierung) oder mit Relationshinweisen (z. B. Konnektoren wie Satzkonjunktionen oder Adverbien) zusammen.

Jede im Text vorkommende Referenz auf die Welt kann die Themaerwartungen auslösen. Bei der Identifizierung eines Teilthemas kann die Tatsache helfen, dass das Thema im Titel erwähnt wird (Themaeinführungshinweise) oder dass bestimmte Wörter im Text wiederholt vorkommen (Themabeibehaltungshinweise). Im Text kann als ein Themahinweis alles verstanden werden, was dazu beiträgt, die Dynamik des Aufbauens, Beibehaltens, Modifizierens und Abbauens von Themaerwartungen zu signalisieren. Die einzelnen Arten der Theamahinweise werden jeweils in selbstständigen Abschnitten vorgestellt, angefangen wird mit den Hinweisen, die das Thema einführen.

5.2.1 Themaeinführungshinweise

Themaeinführungshinweise[81] deuten auf das Aufgreifen und die Fortführung bestimmter Referenzen im weiteren Textverlauf hin, es sind Hinweise wie:

1) Titel und Überschriften – wenn Referenzen auf Welt enthalten sind, können sie als Themaeinführungshinweise betrachtet werden, und auf diese Weise lösen sie thematische Erwartungen des Lesers aus. Ihre Reichweite betrifft den Gesamttext (Titel) oder nur eine textuelle Einheit (wie Kapitelüberschrift).

Der Titel des hier behandelten Romans lautet ‚Unkenrufe'. Im Text kommen Belege vor, in der Bedeutung eines Geräusches, das diese Froschart produziert: *Später, als der Besuch schon über alle Berge war, höre ich sie, vor Uferschilf stehend, über die Unkenrufe hinweg aufs Tonband sprechen:* (GGU, S. 131)

Einige Belege in diesem Text deuten aber nicht auf die Frösche. Das Wort ‚Unkenruf' wird im Deutschen auch übertragen als eine pessimistische Aussage[82] verstanden, diese Bedeutung ist auch in diesem Beispiel präsent: *»Aber verehrte Frau Piątkowska! Was sollen denn diese Unkenrufe?«* (GGU, S. 247) Der Titel des Buches bietet durch seine Doppeldeutigkeit ein starkes Themaerwartungssignal an, weil mehrere Interpretationen möglich sind.

81 Vgl. Hausendorf / Kesselheim 2008, 105ff. Die Beschreibung der Punkte wurde teilweise wörtlich übernommen.

82 Die Definition von ‚Unkenruf' vgl. Duden – Deutsches Universalwörterbuch, 6. Aufl. Mannheim 2006 [CD-ROM].

2) Metakommunikative Hinweise – sie können sich im Text verselbstständigen, es handelt sich um formelhafte Wendungen oder Teiltexte (wie Einführung), mit denen auf das Thema selbstbezüglich referiert werden kann. Im Romantext macht der Autor an sich keine Äußerungen, sondern der Erzähler:

Jetzt könnte ein Briefroman beginnen, dieses knisternde Hin und Her, das bei verstellter Stimme mitteilt, indem es ausspart und fortgesetzt dem Leser mit vielsagenden Lücken zu tun gibt. (GGU, S. 83)

Mit der Wendung *Jetzt könnte ein Briefroman beginnen* gibt der Erzähler einen Hinweis auf den Anfang des Teilthemas in diesem Textabschnitt, in dem es um den Briefwechsel zwischen den Protagonisten geht.

3) Fokus-Hinweise – sie kommen im Laufe der Texte vor und heben die Referenzen auf Welt relativ zur textuellen Umgebung hervor, und die dadurch verursachte Betonung wird zum Auslöser für Themaerwartungen. Es handelt sich um unterschiedliche Mittel dieser Art der Hervorhebung, wie die Hervorhebung durch spezielle Sprachzeichen (z. B. Fokus-Adverbien) oder Herausstellungen. Eine bedeutende Rolle spielt der unbestimmte (‚kataphorische') Artikel, dieser signalisiert die Einführung einer Referenz auf die Welt und ihre Relevanz für das Textthema. Dies bedeutet aber nicht unbedingt, dass das eingeführte Teilthema im Textverlauf weiter enthalten sein wird. Ein Beispiel für ein Einführungssignal für das Teilthema in einem thematischen Strang ist der unbestimmte Artikel bei ‚Einkaufsnetz': *Jedenfalls fand die Witwe in ihrer Umhängetasche ein Einkaufsnetz für die in Zeitungspapier eingeschlagenen Pilze, zu denen die Marktfrau ein Bund Petersilie legte, als Zugabe*. (GGU, S. 13)

Das eingeführte Thema wird im Textverlauf jeweils mit bestimmten Hinweisen unterstützt und beibehalten.

5.2.2 Themabeibehaltungshinweise

Die Themabeibehaltungshinweise[83] signalisieren, dass eine Referenz nicht zum ersten Mal im Text auftaucht; ihre Funktion ist die Beibehaltung einer schon

83 Vgl. Hausendorf / Kesselheim 2008, 115ff. Die Beschreibung der Punkte wurde teilweise wörtlich übernommen.

vorgekommenen Referenz. Durch die Verbindung zu bereits gegebenen referenziellen Hinweisen kommt der Retrospektionscharakter der Themabeibehaltungshinweise zum Ausdruck, und sie stehen in Opposition zu den prospektiven Themaeinführungs-hinweisen. Zu den Themabeibehaltungshinweisen gehören:

1) Anaphorische Themabeibehaltungshinweise – sie werden durch den bestimmten Artikel vertreten. Der anaphorische Artikel zeigt, dass es sich nicht um das erste Vorkommen im Text handelt, sondern dass es um die thematische Beibehaltung des Nomens geht. Ein Beispiel für den anaphorischen Artikel wird in diesem Textabschnitt angeführt: *Schon rede ich, als wäre ich dabeigewesen, von seinem Tweedjackett, von ihrem Einkaufsnetz und verpasse ihm eine Baskenmütze,* [...]*; doch die frühe schon beim Kauf der Steinpilze plazierte Einführung des gehäkelten Erbstücks – die Witwe fand das Netz im Nachlaß ihrer Mutter – ist meine Zutat, wie die vorweggenommene Baskenmütze.* (GGU, S. 16–17) Der bestimmte Artikel bei *die vorweggenommene Baskenmütze* deutet auf die vorher genannte Referenz *eine Baskenmütze*. Derselbe Fall liegt auch bei den Belegen wie *beim Kauf; der Steinpilze; des gehäkelten Erbstücks; die Witwe; das Netz*, die bereits früher im Text erwähnt wurden, vor.

Wenn anaphorische Themabeibehaltungshinweise am Textbeginn vorkommen, ohne dass kataphorische Themaeinführungshinweise vorher erwähnt sind, dann wird damit signalisiert, dass die Szene bereits läuft, wenn die Geschichte beginnt. Dies kann am folgenden Satz illustriert werden:

Der Zufall stellte den Witwer neben die Witwe. (GGU, S. 7)

Es handelt sich um den ersten Satz des Romans, und es sind deswegen keine kataphorischen Hinweise davor zu finden. Die Perspektive der Roman-Protagonisten und der Leser fallen zusammen.

2) Pronominale Themabeibehaltungshinweise – die Pronomen kommen im Text als Rückverweise auf die schon erwähnte Referenzen auf die Welt vor:

Jedenfalls war die Witwe schon zur Stelle, als der Witwer anstieß, stolperte, doch nicht zu Fall kam. Er stellte sich neben sie. (GGU, S. 7)

Diese Pronomen zeigen auf die Substantive des vorstehenden Satzes, *er* auf *der Witwer* und *sie* auf *die Witwe*, es kommt keine neue Information hinzu, das Thema wird durch diese Mittel beibehalten.

3) Themabeibehaltung durch Rekurrenz wird durch die Wiederholung gleicher Wortformen oder -gruppen realisiert, wie bei dem Lexem *Prospekte* in diesem Textabschnitt: *Er gab Reschke einen Stoß Prospekte – »Chatterjees Sightseeing-Tours« – mit auf die Reise: »Für Ihre Freunde in Old Germany! […] Alexander Reschke legte die Prospekte ins Handschuhfach.* (GGU, S. 81)

4) Elliptische Hinweise – die Beibehaltung einer Referenz wird hier durch ihre syntaktische Auslassung im Textverlauf signalisiert. Die Referenz wird als selbstverständlich mitverstanden, wie in einem Fall der Auslassung von Subjekt oder Prädikat. *Hier wird sein!* (GGU, S. 69) In diesem Beispielsatz wird durch das ausgelassene Subjekt ‚der Friedhof' auf den zu gründenden ‚Friedhof' hingewiesen.

Die Referenzen kommen im Text nicht nur wiederholt vor, sondern sie können variieren und zur Entwicklung des Themas beitragen.

5.2.3 Themaentwicklungshinweise

Die Themaentwicklungshinweise[84] zeigen an, dass eine bereits eingeführte Referenz durch andere Referenzen fortgeführt, differenziert und ausgebaut wird und die Themenvariation im Vordergrund steht. Sie sind nicht nur retrospektiv, sondern auch prospektiv, weil sie anzeigen, wie im Text das thematische Potenzial bereits vorkommender Referenzen zu einem thematischen Strang auf- und ausgebaut wird. Die unterschiedlichen Typen der Themaentwicklungshinweise sind:

1) Substitution – ein Referenzausdurck wird durch einen anderen Referenzausdruck mit identischer Referenz (‚Ko-Referenz') ersetzt[85]. Wenn die Variation des ersetzenden Ausdrucks klein ist, dann ist die Grenze zwischen Themabeibehaltungs- und Themaentwicklungshinweisen fließend, wie bei den Substitutionen durch

84 Vgl. Hausendorf / Kesselheim 2008, 120ff. Die Beschreibung der wurde Punkte teilweise wörtlich übernommen.
85 Vgl. Hausendorf / Kesselheim 2008, 121.

Synonyme. Ist die Variation dagegen groß, steht die Themaentwicklung im Vordergrund. Konkrete Realisierungen können Beziehungen zum ursprünglichen Referenzausdruck wie Hyponymie, Hyperonymie oder eine losere Art lexikalischer Beziehung sein.

Ein Beispiel für die Substitution durch ein Hyperonym bietet dieser Textabschnitt: *Wenn nicht die Blumen, darf ich, bitte, dann den Gegenstand unseres gerade begonnenen Gesprächs, einige Steinpilze, diesen hier, den, den und noch den, auswählen und Ihnen zum Geschenk machen?* […] *Sicher ist: vorm Kauf fotografierte er die Pilze und nannte die Firmenmarke seiner Kamera japanisch*. (GGU, S. 12) Das Lexem ‚Pilze' weist auf die vorher im Text erwähnten ‚Steinpilze' zurück.

2) Lexikalische Themaentwicklungshinweise deuten auf die semantischen Beziehungen zwischen den Referenzen hin, die nur im jeweiligen Text deutlich werden, wie z. B.:

- Hypero- und Hyponymie (Ober- und Unterbegriff) bzw. Holo- und Meronymie (Ganzes- und Teilbegriff) weisen auf einen Übergang von einem zum anderen erwähnten Ausdruck bzw. umgekehrt hin. In diesem Beispiel wird das Einkaufsnetz durch den Oberbegriff ‚Netz' wiederholt: *Jedenfalls fand die Witwe in ihrer Umhängetasche ein Einkaufsnetz für die in Zeitungspapier eingeschlagenen Pilze, zu denen die Marktfrau ein Bund Petersilie legte, als Zugabe. Er wollte das Netz tragen*. (GGU, S. 13)

- Antonymie (Gegensatz-Verhältnis) – die Gegensatz-Relationen sind oft gegenstands- und kontextabhängig. In diesem Beispiel stehen gegeneinander ‚(auf Hochglanz) poliert' und ‚rostig': *»Auf grauem Sockel ein schwarzer, auf Hochglanz polierter Granit. Die geräumige Grabstelle faßt ein rostiger Eisenzaun ein.* (GGU, S. 199) Der Ausgangspunkt dieser Relation ist die Opposition von ‚im guten Zustand' – ‚verkommen'. Die antonyme Beziehung dieser zwei Begriffe ist stark kontextgebunden.

- Metaphorik und Metonymie, Homonymie – stellen die Beziehungen der Bedeutungsübertragung her. Der Beispieltextabschnitt zeigt den Übergang von wörtlicher zu übertragener Bedeutung an: *»Glaub mir, Alexandra, wie der Raps zu früh blüht, rufen Rotbauchunke und Gelbbauchunke. Sie wollen uns etwas sagen ...«* [...]

»Bist selber Unke!« (GGU, S. 127) Im Deutschen wird die Person, die Unglück prophezeit[86], übertragen als *Unke* bezeichnet. Reschke wird als Unke bezeichnet, weil er pessimistische Aussagen macht.

3) semantische Themaentwicklungshinweise – gehen von den semantischen Beziehungen (die den lexikalischen Hinweisen nahe stehen) aus. Sie können von zwei eng verbundenen Seiten betrachtet werden:

- Isotopiehinweise – diese Hinweise bestehen in der Rekurrenz semantischer Merkmale. Im Unterschied zu den standardisierten lexikalischen Relationen geht es um die textbezogene Aktivierung der semantischen Merkmale. Diese Merkmale ergeben sich jeweils aktuell in Differenz zu anderen Lexemen. Innerhalb einer Isotopie werden unterschiedliche Referenzausdrücke in einem Text durch die Wiederkehr eines semantischen Merkmals zusammengeführt und die Zusammengehörigkeit zu einem Referenzbereich festgelegt.

Bei der Suche nach dieser Art der Hinweise spielen die Konnotation und die Thematik des Textes eine bedeutende Rolle. Als ein Beispiel wird die Beziehung zwischen ‚Einkaufsnetz' und ‚Erbstück' betrachtet: *Das Einkaufsnetz ist keine Erfindung.* [...] *doch die frühe schon beim Kauf der Steinpilze plazierte Einführung des gehäkelten Erbstücks – die Witwe fand das Netz im Nachlaß ihrer Mutter – ist meine Zutat, […] waren ihm nun die Einkaufsnetze der Witwe – sie erbte nicht nur das eine, sondern ein halbes Dutzend – Zeugnisse vergangener Kultur, verdrängt von häßlichen Wachstuchtaschen und radikal entwertet durch den Plastikbeutel*. (GGU, S. 16–17)

Die Belege *Das Einkaufsnetz* und *des gehäkelten Erbstücks* werden durch die Merkmale ‚gehäkelt', ‚geerbt' und ‚ein zum Transport von Einkauf dienender Gebrauchsgegenstand' verbunden. Der Kontext erschließt den Zusammenhang durch die Tatsache, dass die Witwe das Einkaufsnetz geerbt hat. Diese Beziehungen können nicht formalisiert werden, bei jedem Einzelfall muss aufmerksam vorgegangen werden.

86 Zur Definition von ‚unken' vgl. Duden – Deutsches Universalwörterbuch, 6. Aufl. Mannheim 2006 [CD-ROM].

- Rahmenhinweise – sie beziehen sich auf die Vertrautheit mit einem Handlungsrahmen (‚frame') oder einem Handlungsablauf (‚script').[87] Diese Hinweise werden oft durch lexikalische Hinweise und Isotopiehinweise begleitet. Sie sind mit der Aktivierung von Schemata der Wissensrepräsentationen verbunden, und es werden Themaerwartungen eingeführt, z. B. bei dem Lexem ‚Friedhof':

*Er fand am Rand des **Friedhofs** zwei schiefstehende **Steine**, später zwei weitere, gänzlich verkrautet, und hatte Mühe, ihnen irgendwas abzulesen. Mit weit zurückliegenden Sterbedaten – Anfang der zwanziger bis Mitte der vierziger Jahre – und mit Inschriften über den Namen – »Hier ruht in Gott«, »Der Tod ist das Tor zum Leben« oder »Hier liegt unsere liebe Mutti und Omi« – erinnerten sie an die Vorvergangenheit der **Friedhofsanlage**. Reschke notiert: »Auch diese **Steine** aus üblichem Material: Diabas und schwarzschwedischer Granit.« Für ein Weilchen lasse ich ihn bei den übriggebliebenen Steinen. Frau Piątkowska wird inzwischen den Strauß Astern am **Grab** ihrer Eltern in eine Vase gestellt haben. Dieser **Doppelgrabstelle** sage ich nach, daß sie, buchsbaumumrandet, weniger überwuchert ist als die benachbarten **Grabstellen**.* [...] *Auf allen **Feldern** kann ich **Allerseelen-Betrieb** beobachten: hier und da bezeugen Windlichter an **Grabstellen** Besuch, der wieder gegangen ist.* (GGU, S. 22–23)

Das Verhältnis zwischen den Wörtern in diesem Rahmen kann logisch-begrifflich (*Friedhof – Steine – Grab – Doppelgrabstelle – Grabstellen – Feld*), kulturell (*Friedhof – Allerseelen-Betrieb*) auch lexikalisch (*Friedhof – Friedhofsanlage*) begründet werden. Es werden also unterschiedlich umfangreiche Referenzbereiche aufgrund der genannten Wortschatzgruppierungsarten (Wortfamilien, Sachgruppen, Wortfelder) zusammengestellt.

Bei der Unterscheidung zwischen den Isotopie- und Rahmenhinweisen hängt die Bestimmung von der Interpretation der jeweiligen Textstelle ab. Als Beispiel wird hier das Lexem ‚Sekretariat' angeführt:

*Er richtete in seiner Junggesellenwohnung ein **Sekretariat** ein* [...] (GGU, S. 132) Aufgrund des Weltwissens ist festgelegt, dass eine Gesellschaft ein Sekretariat betreibt, und es könnte sich hier um einen Rahmenhinweis handeln. Da es sich in

87 Siehe Hausendorf / Kesselheim 2008, 132.

diesem Roman um eine konkrete Gesellschaft mit einem konkreten Sekretariat handelt, wird dem Kontext größerer Wert zugeteilt, und dieser Fall kann als Isotopiehinweis bewertet werden.

Die Isotopie- und Rahmenhinweise ergänzen sich gegenseitig, und es kommt zu Überlappungen, deswegen wird in der späteren Untersuchung des Teilthemas die Unterscheidung dieser zwei Arten der Hinweise nicht durchgeführt und die Belege werden nur als semantische Themaentwicklungshinweise bezeichnet.

Wenn ein Thema im Textverlauf nicht mehr relevant ist, wird es durch entsprechende Mittel abgeschlossen.

5.2.4 Themaabschlusshinweise

Die Themaabschlusshinweise[88] zeigen an, dass bestimmte Referenzen im Text keine Rolle mehr spielen; entweder wird die Gesamtthematik eines Textes oder es werden einzelne thematische Stränge abgeschlossen. Das Thema wird also beendet, wenn es im Text keine Themabeibehaltungshinweise mehr gibt, oder wenn das Ende des Themas metakommunikativ signalisiert wird. Oft decken sich die Themaabschlusshinweise mit dem Ende der textuellen Einheiten, und eventuell kann durch eine Abschlussformel das Ende der Geschichte und gleichzeitig das Ende weiterer Referenzen signalisiert werden. Der Abschluss des Hauptthemas ist im hier behandelten Romantext an die Textgrenze gebunden.

Im Text können bestimmte abgeschlossene Themen wieder an Relevanz gewinnen und erneut eingeführt werden.

5.2.5 Themawiedereinführungshinweise

Eine bereits eingeführte und evtl. entwickelte und abgeschlossene oder in den Hintergrund getretene Referenz wird erneut durch die Themawiedereinführungshinweise auffällig gemacht[89].

88 Vgl. Hausendorf / Kesselheim 2008, 134ff. Die Beschreibung der Punkte wurde teilweise wörtlich übernommen.

89 Vgl. Hausendorf / Kesselheim 2008, 135ff. Die Beschreibung der Punkte wurde teilweise wörtlich übernommen.

1) Metakommunikative Hinweise – um die Bezugnahme deutlich zu machen, werden Thematisierungsausdrücke und -formeln verwendet. Im Roman-Text spricht nicht der Autor, sondern der Erzähler:

Aber er will gesehen haben, wie ich eine ausgewachsene Kröte, nein, Unke, Rotbauchunke geschluckt, ohne zu würgen, verschluckt, runtergeschluckt habe, rein und weg, ohne Wiederkehr. (GGU, S. 43) […] *Also gut, ich habe als Schüler auf Wunsch Kröten geschluckt.* (GGU, S. 60)

Durch die im Vor-Vorfeld stehende Rethematisierungsformel *Also gut* referiert der Erzähler auf einen wesentlich weiter vorne im Text vorkommenden Abschnitt, in dem über das Schlucken von Kröten erzählt wird.

2) Rhematische Pronominalisierung bedeutet eine Kombination von Themaeinführung und Themabeibehaltung. Es wird signalisiert, dass die Referenz nicht zum ersten Mal vorkommt, und sie wird auffällig eingeführt mit besonderer Relevanz der wiedereingeführten Referenz:

Nur kurz flog ihn der Geruch der Radaune an. »Nein, Alexandra, die hat schon immer, jedenfalls meine Schulzeit lang, so gestunken.« (GGU, S. 45)

Die rhematische Pronominalisierung durch ‚die' bei *die hat schon immer* [...] *gestunken* führt die Referenz erneut auffällig ein.

3 Demonstrativ-Artikel – auffällige Rethematisierung mit dem Demonstrativ-Artikel ‚dieser/diese/dieses':

Reschke hält fest, daß er ***das Gelände zwischen Brentau und Matern*** *anfangs als geeignet für einen Waldfriedhof gesehen habe, zumal »die hochstämmigen Buchen gut Abstand halten und dem Projekt einen natürlichen Rahmen leihen. Nur wenige Exemplare wird man fällen müssen. Das viele Niederholz hingegen. Denn nicht versteckt, wohl aber geborgen unterm Blätterdach sollen die heimgekehrten Toten ihre Ruhe finden.« Gegen* ***diese ideale Friedhofslage*** *sprach der zu nah gelegene Flughafen von Gdańsk, dessen Landebahnen dort, wo früher die Gehöfte des Dorfes Bissau ihre leicht gehügelten Äcker um sich versammelt hatten, planen Raum einnahmen.* (GGU, S. 64)

Die Verwendung des Demonstrativ-Artikels bei der Wortgruppe *diese ideale Friedhofslage* signalisiert die Verbindung zu dem vorher genannten *Gelände zwischen Brentau und Matern*.

4) Herausstellungen – die Links- und Rechtsversetzungen nützen die Topologie für Fokus-Hinweise, dieses Beispiel zeigt die Herausstellung nach rechts:

Die beiden Schulfreundinnen riefen: »Können wir Ihnen wärmstens empfehlen, dieses Vergnügen! (GGU, S. 60) In diesem Abschnitt wird die früher im Text erwähnte Rikschafahrt durch die Herausstellung von *dieses Vergnügen* auffällig hervorgehoben.

Es wurden Themahinweise vorgestellt, so wie sie im Textverlauf vorkommen können. Sie werden auch bei der Beschreibung eines Teilthemas, das dem Hauptthema untergeordnet ist, verwendet.

5.3 Teilthemaentwicklung im Text

Die vorgestellten Erkenntnisse zum Textthema werden auf ein ausgewähltes Phänomen angewendet, um die Vorgehensweise für die weitere Untersuchung zu entwerfen. Das Verständnis des Textthemas geht hier textsemantisch von dem Textinhalt aus. Während der Untersuchung werden die Themahinweise, die auf die thematische Progression hindeuten, beobachtet.

5.3.1 Analyse an einem Modellfall

Durch die Bearbeitung und Untersuchung des Gesamttextes mithilfe der Wortformenliste (die im Unterkapitel „4.1.1 Wortformenliste“ beschrieben wurde) wird die ‚Friedhofsgesellschaft‘ als das zu untersuchende Teilthema im Gesamttext bewertet. Im Textverlauf sind auch andere Nebenhandlungen vorzufinden, die diese Handlung begleiten[90]. Es sind entweder Nebenhandlungen, die im Laufe des ganzen Textes auftauchen, wie zum Beispiel:

a) räumliche Hinweise darauf, dass sich die Geschichte überwiegend in der Stadt Gdańsk abspielt

b) gemeinsame Vergangenheit von Reschke und Erzähler

c) das Rikscha-Unternehmen von Chatterjee

90 Die aufgeführten Beispiele der Nebenhandlungen dienen zur Illustration, sie stellen keine vollständige Auflistung aller im Text vorkommenden Fälle dar.

– oder es sind Abschnitte, die eine Situation oder einen Gegenstand betreffen und in der Regel die Grenze eines Kapitels nicht überschreiten oder im benachbarten Kapitel vorkommen, jedoch nicht im Verlauf des Gesamttextes zu verzeichnen sind. Zum Beispiel:

d) Engel, an dessen Vergoldung Piątkowska arbeitet

e) Videoaufnahme der ersten Beerdigungen

f) Tonaufnahmen der Sprache von Brakup

Die in der Erzählung präsenten Nebenhandlungen bilden eigene Isotopien, sog. Isotopiestränge (bzw. Isotopieebenen), und sie sind dem Hauptthema untergeordnet (vgl. THIEL 1996, 61ff).

Ein ausgewählter Vertreter der Gruppe der kleineren Isotopiestränge wird näher betrachtet, und als das zu untersuchende Phänomen wird die Isotopie „ENGEL“[91] angeführt. Der untersuchte Textabschnitt wird hier gekürzt angeführt:

> *Dort hatte die Piątkowska gerade* ***einem spätgotisch knienden Engel***[1] *im Bereich des rechten Flügels die Mastixschicht aufgetragen und ihren Alexander gebeten, Kaffeewasser aufzusetzen, als der Freund eintraf und von Erna Brakups Krankenlager berichtete.* (239) *Mit* ***dem Engel***[2] *mußte der Kaffee warten.* (240)
>
> […] Reschke und Piątkowska besuchen die kranke Brakup …]
>
> *Dort kniete spätgotisch* ***der Engel***[3]*, dessen Mastixgrundierung auf Kreidegrund weiteren Auftrag von Blattgold erwartete.* (242) […] ***Der Engel***[4] *kniete auf dem Küchentisch:* [**er**[5] war] *einen knappen Meter hoch.* (242)
>
> […] Wróbel (ein Aufsichtsratmitglied, das die ursprüngliche Versöhnungsidee vertritt), Reschke und Piątkowska reden über die Situation in der Gesellschaft …]
>
> *Sie nahmen sich die Mitglieder einzeln vor, verglichen witzig Bieroński mit Karau, sprachen kurz über das leidige Projekt »Bungagolf«, erwogen eine gemeinsame Frühjahrsreise nach Schlesien, um dort die neuentstandenen Versöhnungsfriedhöfe zu besuchen, nannten zusätzlich Allenstein und Stolp als Reiseziele, wollten sogar in Bromberg einen letzten Versuch wagen, lobten die Suppenküchen einiger Seniorenheime und bewerteten ihre Tätigkeit – vom Umbettungsgeschäft abgesehen – als sinnvoll, weil versöhnend, da sprach Jerzy*

91 Die Zahlen nach den Sätzen geben die Seitenzahl im gedruckten Buch ‚Unkenrufe‘ von Günter Grass an, um anzuzeigen, über welchen Textabschnitt sich diese Isotopie erstreckt. Die hochgestellten Zahlen stellen die Nummerierung der Belege in dieser Untersuchung dar.

Wróbel, während die Piątkowska das Blattgold auf dem ***Engelsflügel***[6] *nun mit weichem Flachpinsel verstrich, wie beiläufig das Wort Rücktritt aus; und schon während der nächsten Aufsichtsratssitzung hat er seinen Rücktritt erklärt.* (243)

[… bei der nächsten Sitzung werden neue Projekte vorgestellt, wie die mehrsprachigen Straßenschilder; Wróbel tritt ab, Reschke und Piątkowska machen sich Gedanken, wie es mit der Gesellschaft weiter geht …]

Oder waren sie auf Gedankenflucht: heimwärts in Alexandras Küche ***zum knienden Engel***[7]*?* (245)

[… bei der Aufsichtsrat-Sitzung betont Reschke den ursprünglichen Versöhnungsgedanken, und unter den Umständen der geschäftlichen Entwicklung der Gesellschaft meldet er seinen und Piątkowskas Rücktritt an, beide bleiben jedoch noch als Ehrenvorsitzende; daneben wird auch von dem Scheitern der Litauischen Komponente berichtet. Neue Aufsichtsratsmitglieder werden eingesetzt, und bei der Sitzung werden neue Projekte behandelt. Schließlich bittet Piątkowska Reschke die Sitzung früher zu verlassen …]

Sie sagte: »Du weißt, ***Engel***[8] *wartet auf uns in Küche.«* (250)

[… die Gesellschaft hat neues Gelände gepachtet; im nächsten Abschnitt wird das Vergolden der Engelfigur beschrieben …]

Der meterhoch kniende Engel[9] *auf dem mit Zeitungspapier abgedeckten Küchentisch.* (251) […] *Wie sakrale Gegenstände feiert er ihr Werkzeug und nennt das Vergolderkissen, auf dem sie das Blattgold in passende Quadrate schneidet, »ihren wie eine Palette geführten Altar, von dem sie mit einer aus Buchsbaum geschnitzten Pinzette feierlich langsam hauchzarte Blättchen abhebt, um sie mit dem Anstauchpinsel aus weichem Kamelhaar dem Kreidegrund auf dem Altholz* ***des Engels***[10] *anzudrücken.* (251) […] *Den nur noch Ehrenvorsitzenden der Friedhofsgesellschaft wurde* ***die windstill gehaltene Küche, in der ein spätgotisch kniender Engel***[11] *immer mehr Blattgold annahm, zum zentralen Ort.* (252)

[… in der Zwischenzeit besucht das Paar die kranke Brakup, insgesamt gehen sie kaum noch aus …]

Er nannte ***den Engel***[12] ***»eine anonyme, wahrscheinlich süddeutsche, wenn nicht böhmische Arbeit***[13]*«.* (253) […] ***Der kniende, aus Lindenholz geschnitzte Engel***[14] *blies eine Posaune: so konnte* ***er***[15] *dem Jüngsten Gericht zugeordnet werden.* (253) *»Ursprünglich waren es viele, wird es ein Chor* ***Engel***[16] *gewesen sein, der, gänzlich vergoldet, durch Posaunenschall Grüfte gesprengt, Gräber geöffnet, Beinhäuser aufgeschlossen hat, auf daß in Erfüllung ging, was ich kürzlich wieder einer Bodenplatte im Mittelschiff von Sankt Trinitatis abgelesen habe: ›Nach vollbrachter Mueh und Jammer / Ruh ich hier in meine Kammer / bis ich eins werd Auferstehen / Und zur ewigen Freid*

eingehen.‹ (253) *Und dazu gab* ***Alexandras Engel***[17] *das Signal …«* (253) *Sie hatte* ***die holzgeschnitzte Figur***[18] *aus anderer Werkstatt übernommen.* (253) […] ***Der auf linkem Knie kniende Engel***[19] *muß in* ***seinem***[20] *geflickten Zustand erbärmlich ausgesehen haben:* [**er**[21] war] ***ein Veteran***[22] ***wechselnder Zeitläufte****.* (253) *Selbst nach dem komplizierten Auftragen der acht leimgebundenen Kreideschichten war* ***ihm***[23]*, wenngleich die Vergolderin keine Feinheiten der spätgotischen Faltenwürfe zuschlämmte, viel von* ***seiner***[24] *möglichen Schönheit vergangen.* (254) *Doch als* ***der kniende Engel***[25] *zusehends zu Gold kam, mit beiden Flügeln gülden glänzte, als Alexandra, nachdem sie das zwischen Weiß- und Rotgold legierte Material der Mastixschicht überm Kreidegrund aufgetragen hatte, schließlich den Goldbelag vom Posaunentrichter bis zu den Zehenspitzen mit dem Achatstein zu polieren begann, entstand* ***die Figur***[26] *in* ***ihrer***[27] *von anonymer Hand gewollten Schönheit aufs neue.* (254) *Der vorher grämlich wirkende Ausdruck* ***des blasenden, langgelockten, eher einen jungen Mann als eine Jungfrau im Faltenwurf verbergenden Engels***[28] *gewann jene herbe Anmut, die, wie Reschke sagt, »frühe* ***Riemenschneiderengel***[29] *auszeichnet …«.* (254) *Dennoch hat er* ***die wiederbelebte Arbeit***[30] *künstlerisch nicht allzu hoch eingeschätzt: »Mit anderen Figurationen wird* ***der Engel***[31] ***Beiwerk***[32] ***eines Altars*** *gewesen sein, als dessen zentrales Motiv ich die Auferstehung vermute.* (254) *Erstaunlich, wie* ***das heruntergekommene Stück***[33] *unter Alexandras Händen gewonnen hat.* (254) *Immer wieder – nun beim Polieren der höher gelegenen Partien – verspricht sie sich und mir: ›Wirst sehen,* [**er**[34]] *wird sein wie neugeboren.‹« (254) Als Alexandra Piątkowskas Küche ganz und gar vom hochglanzvergoldeten* ***Auferstehungsengel***[35] *bewohnt war, brachte der städtische Angestellte Jerzy Wróbel die Nachricht vom Tod der Erna Brakup.* (254)

Diese Szene spielt sich im sechsten Kapitel ab, die Friedhofsgesellschaft prosperiert, weil es aber immer mehr um Geschäfte geht, tritt das Aufsichtsratmitglied Erna Brakup aus dem Aufsichtsrat aus. Auch die Gründer der Idee, Reschke und Piątkowska, ziehen sich allmählich zurück. Der beobachtete Referenzbereich des hier untersuchten Textabschnitts ist der Engel, den Piątkowska zu Hause restauriert, und dieses Teilthema wird von einer Nebenhandlung ‚die Besuche bei der kranken Brakup' begleitet. Die Musteruntersuchung wird nur an dem Teilthema ‚Engel' durchgeführt.

Im Gesamttext kommt das Wort *Engel* noch an zwei anderen Stellen vor, diese Vorkommen gehören aber nicht zu diesem Referenzbereich. Der erste Beleg von *Engel* findet sich im ersten Kapitel, in einer Situation, in der sich das Paar gegenseitig im Gespräch vorstellt. Die Vergolderin Piątkowska spricht davon, dass sie u. a. auch *Barockengel* (GGU, S. 27) vergoldet, der zweite Beleg ist *Engelsgrube* (GGU, S. 110),

als Name einer Lübecker Straße. Es muss aber hinzugefügt werden, dass sich das Wort *Engel* bei dem Wort *Engelsgrube* auf England bezieht, nicht auf Engel als Boten Gottes.[92] Also referiert keiner von diesen zwei Belegen auf den hier besprochenen Engel, der in der Küche Piątkowskas vergoldet wird. Im Gesamttext gibt es kein von dem Substantiv *Engel* abgeleitetes Adjektiv. Die im Text vertretenen Adjektive ‚englisch' beziehen sich auf die englische Sprache, nicht auf den *Engel*.

Das Ziel dieser Untersuchung ist es, die Beziehungen innerhalb eines eher kleineren Referenzbereiches zu betrachten und die Vorgehensweise für die Untersuchung der Teilthemaentwicklung im Gesamttext zu entwerfen.

Die Aufteilung der Beschreibungsphasen ergibt sich aus den bei der Untersuchung durchgeführten Schritten: Zuerst wird die von dem Computer angebotene Suchfunktion genutzt, um im Korpus die Ausdehnung des Referenzbereiches im Text mindestens grob zu erfassen, d. h., es wurde nach dem Schlüsselwort ‚engel'[93] gesucht (Punkt a), und danach wird im Text nach anderen relevanten Vorkommen recherchiert (Punkt b). Um den Zusammenhang zwischen den einzelnen Belegen zu prüfen, werden die Kohäsions- und Kohärenzbeziehungen betrachtet (Punkt c), denn diese Art der Beziehungen zwischen den Wörtern hilft, die Themahinweise (Punkt d) im Text zu entdecken:

a) die Vorkommen des Schlüsselwortes ‚Engel', auch als Teil einer Wortbildung: Bei der Suche wird überwiegend von der Suche nach Zeichenketten ausgegangen, dabei können auch die Beziehungen aufgrund der Zugehörigkeit zu der jeweiligen Wortfamilie unterstützt werden. Der Zeichen-String ‚engel' kommt auch in Wörtern wie *Unkengeläut* oder *offengelegt* vor, aber diese Fälle haben mit dem ‚Engel' nichts gemeinsam, deswegen werden sie beim Sammeln der Belege nicht beachtet.

Die Zugehörigkeit zu diesem Referenzbereich stützt sich hier auf das Vorkommen des Lexems ‚Engel', deswegen sind die einfachen Wiederholungen dieses Wortes besonders auffällig:

> *einem spätgotisch knienden Engel*[1]*; Mit dem Engel*[2]*; der Engel*[3]*; Der Engel*[4]*; zum knienden Engel*[7]*; Engel*[8]*; Der meterhoch kniende Engel*[9]*; des Engels*[10]*; die windstill gehaltene Küche, in der ein spätgotisch kniender Engel*[11]*; den*

92 Vgl. http://de.wikipedia.org/wiki/Engelsgrube, zit.: 25. 1. 2011.

93 Wenn nicht anders angegeben, spielt die Klein- oder Großschreibung bei der Suche keine Rolle.

Engel[12]*; Der kniende, aus Lindenholz geschnitzte Engel*[14]*; Alexandras Engel*[17]*; ein Chor Engel*[16]*; Der auf linkem Knie kniende Engel*[19]*; der kniende Engel*[25]*; des blasenden, langgelockten, eher einen jungen Mann als eine Jungfrau im Faltenwurf verbergenden Engels*[28]*; der Engel*[31]

Die Verwendung des Lexems ‚Engel' innerhalb der Wortbildungen wurde nur bei Determinativkomposita gefunden:

auf dem Engelsflügel[6]*; Riemenschneiderengel*[29]*; Auferstehungsengel*[35]

Bei dem Beleg Nr. 6 steht der Engel in der Rolle des Bestimmungswortes: *Engelsflügel*. In den restlichen Determinativkomposita wird ‚Engel' als Grundwort *Riemenschneiderengel*[29] und *Auferstehungsengel*[35] verwendet.

b) weitere unterschiedliche Lexeme, die im Text auf den in der Küche knienden Engel hinweisen:

Eine andere Art der Wiederaufnahme des Wortes ‚Engel' spiegelt sich auch in der Verwendung von Pronominalisierungen:

Der kniende, aus Lindenholz geschnitzte Engel[14] *blies eine Posaune: so konnte* ***er***[15] *dem Jüngsten Gericht zugeordnet werden.*

Der auf linkem Knie kniende Engel[19] *muß in* ***seinem***[20] *geflickten Zustand erbärmlich ausgesehen haben:* [**er**[21] war] ***ein Veteran***[22] ***wechselnder Zeitläufte****.*

Selbst nach dem komplizierten Auftragen der acht leimgebundenen Kreideschichten war ***ihm***[23]*, wenngleich die Vergolderin keine Feinheiten der spätgotischen Faltenwürfe zuschlämmte, viel von* ***seiner***[24] *möglichen Schönheit vergangen.*

Doch als der kniende Engel[25] *zusehends zu Gold kam, mit beiden Flügeln gülden glänzte,* [...] *schließlich den Goldbelag vom Posaunentrichter bis zu den Zehenspitzen mit dem Achatstein zu polieren begann, entstand die Figur*[26] *in* ***ihrer***[27] *von anonymer Hand gewollten Schönheit aufs neue.*

Einerseits kommen hier die Personalpronomen *er*[15] und *ihm*[23] vor, andererseits sind auch die Possessivpronomina *seinem*[20] und *seiner*[24] vertreten. Diese Pronomina beziehen sich auf ein maskulines Substantiv, das in diesem Fall eben der Engel ist. Der Beleg *ihrer*[27] bezieht sich auf *die Figur*[26], die im Text referenzidentisch mit ‚Engel' ist. Keines dieser auf den Engel hinweisenden Pronomen steht aber in einem Satz selbstständig.

Auf das hier behandelte Beispiel beziehen sich Substitutionen durch andere (substantivische) Lexeme, die in dasselbe Wortfeld bzw. den Referenzbereich mit dem Lexem ‚Engel' einzuordnen sind:

> *eine anonyme, wahrscheinlich süddeutsche, wenn nicht böhmische **Arbeit**[13]; die holzgeschnitzte **Figur**[18]; ein **Veteran**[22] wechselnder Zeitläufte; die wiederbelebte **Arbeit**[30]; **Beiwerk**[32] eines Altars; das heruntergekommene **Stück**[33]*

Die semantischen Merkmale des untersuchten Lexems ‚Engel' ergeben sich aus den unterschiedlichen Vorkommensformen und dem Kontext wie: ‚eine manuell aus Holz geschnitzte Statue menschlicher Gestalt mit Flügeln, die alt ist und restauriert wird'. Außerdem ist es die Funktion einer solchen Statue, einen Altar als Schmuck zu ergänzen. Damit die hier genannten Belege zu demselben Referenzbereich (bzw. Sachgruppe) gehören können, müssen sie sich an ihm semantisch beteiligen. Diese Beziehungen werden von dem Kontext unterstützt.

Ähnlich wie der Kontext können auch andere Teile der Mehrwortverbindungen die Bedeutung der jeweiligen Belege präzisieren. Anhand der ausgewählten Belege kann der Engel näher identifiziert werden, damit die Referenzidentität deutlicher wird, wie z. B.:

> *- der Engel[3], dessen Mastixgrundierung auf Kreidegrund weiteren Auftrag von Blattgold erwartete*
> *- Der meterhoch kniende Engel[9]*
> *- Altholz des Engels[10]*
> *- Der kniende, aus Lindenholz geschnitzte Engel[14]*

Aus den unterschiedlichen Belegen kann die Charakteristik des konkreten Engels aufgebaut werden. Es handelt sich also um einen ein Meter hohen, holzgeschnitzten Engel, der restauriert wird. Durch das Beobachten der einzelnen Bedeutungsmerkmale wird die Verbindung zu dem Substituendum ‚Engel' deutlicher:

Beleg Nr. 13 (*eine anonyme, wahrscheinlich süddeutsche, wenn nicht böhmische Arbeit*[13]) – eine allgemeinere Bezeichnung der holzgeschnitzten Engelfigur mit dem Oberbegriff ‚Arbeit' als ein durch Betätigung entstandenes Werk[94]

Beleg Nr. 18 (*die holzgeschnitzte **Figur***[18]) – die Gestalt des Engels wird mit dem referenzidentischen Nomen ‚Figur' bezeichnet

94 Zur Definition von ‚Arbeit' vgl. Duden – Deutsches Universalwörterbuch, 6. Aufl. Mannheim 2006 [CD-ROM].

Beleg Nr. 22 (*ein **Veteran**[22] wechselnder Zeitläufte*) – die Engelfigur ist schon alt und wird restauriert, deswegen wird sie übertragen beschrieben als ‚Veteran', als jemand, der altgedient ist[95], mit dem unbestimmten Artikel wird die Zugehörigkeit zu einer Gruppe signalisiert

Beleg Nr. 30 (*die wiederbelebte **Arbeit***[30]) – ähnlich wie bei Beleg Nr. 13

Beleg Nr. 32 (***Beiwerk** eines Altars*[32]) – Funktion der Engelfigur ist es, einen Altar als ‚Beiwerk' zu schmücken

Beleg Nr. 33 (*das heruntergekommene Stück*[33]) – durch ‚Stück' wird der Engel mit einem Hyperonym bezeichnet bzw. als ein einzelner Gegenstand umschrieben.

Der Zusammenhang dieser unterschiedlichen Lexeme mit konkreten Lexem ‚Engel' ist eng an den jeweiligen Kontext gebunden. Es handelt sich oft nicht um Synonyme, sondern um referenzidentische Lexeme, die zum gleichen Referenzbereich gehören. Die beschriebenen Verbindungen können textkonstitutive Funktion annehmen.

c) bei den aufgefundenen Belegen werden die satzverknüpfenden Beziehungen betrachtet. Die Arten der Wortschatzgruppierung können bei der Suche nach Belegen kombiniert werden, d. h., die Sätze werden mit Ausdrücken verbunden, die auf den Prinzipien der Rekurrenz (z. B. Wortfamilie) oder Substitution (Sachgruppen oder Wortfelder) basieren:

- Repetition identischer Lexeme, Beispiel:

> *Dort hatte die Piątkowska gerade **einem spätgotisch knienden Engel**[1] im Bereich des rechten Flügels die Mastixschicht aufgetragen und ihren Alexander gebeten, Kaffeewasser aufzusetzen, als der Freund eintraf und von Erna Brakups Krankenlager berichtete.* (239) *Mit **dem Engel**[2] mußte der Kaffee warten.* (240)

Die Zugehörigkeit der einzelnen Lexeme ‚Engel' zueinander wird dadurch betont, dass, nachdem der erste Beleg mit dem unbestimmten Artikel als etwas Neues eingeführt wird, weitere Belege (Nr. 2, 3, 4, 6, 7, 9, 10, 12, 14, 19, 25, 28 und 31) mit dem bestimmten Artikel auf den am Anfang erwähnten Beleg Nr. 1 verweisen.

95 Zur Definition von ‚Veteran' vgl. Duden – Deutsches Universalwörterbuch, 6. Aufl. Mannheim 2006 [CD-ROM].

Auch andere Referenzen auf den Engel kommen wiederholt vor: Das zweite Beispiel ist die Wiederholung von ‚Figur', die im Text auf den ‚Engel' verweist: *die holzgeschnitzte Figur*[18] und *die Figur*[26]; *eine anonyme, wahrscheinlich süddeutsche, wenn nicht böhmische Arbeit*[13] und *die wiederbelebte Arbeit*[30] und die Personalpronomina: *er*[5], *er*[15], *er*[21] und *ihm*[23].

In dem untersuchten Textabschnitt kommt noch einmal eine Verwendung des Lexems ‚Engel' mit dem unbestimmten Artikel vor: *Den nur noch Ehrenvorsitzenden der Friedhofsgesellschaft wurde die windstill gehaltene Küche, in der ein spätgotisch kniender* ***Engel***[11] *immer mehr Blattgold annahm, zum zentralen Ort*. Dieses Lexem bezieht sich auf die Beschreibung der Küche, der unbestimmte Artikel deutet nicht auf die erstmalige Verwendung im Text hin.

In einigen Fällen wird auch die Verwendung ohne Artikel realisiert. Der Beleg ‚*Engel*[8] *wartet auf uns in Küche*' stammt aus dem Textabschnitt der direkten Rede von Piątkowska, die oft auf die Verwendung von Artikeln im Deutschen verzichtet. In der Verbindung *ein Chor Engel* steht ‚Engel' als substantivisches Attribut im Plural, die Beziehung zu dem Referenzbereich ist nur die Wiederholung des Lexems ‚Engel'. Bei *Alexandras Engel* wird die Zugehörigkeit durch den Eigennamen ausgedrückt und nach dem Eigennamen ohne Artikel gebracht. In diesen Fällen wird bei der Beurteilung des Zusammenhangs von dem wiederholten Vorkommen des jeweiligen Wortes ausgegangen.

Außer der Repetition der identischen Lexeme können im Text die Topiks mit Wortbildungselementen betrachtet werden.

- Topiks mit Wortbildungselementen

Die Wiederholung von ‚Engel' wird auch innerhalb von Wortbildungen betrachtet. Der Beleg *Engelsflügel*[16] steht im Text mit dem bestimmten Artikel. Die Paraphrase zeigt im Allgemeinen, dass es sich um einen Flügel eines Engels handelt, und aufgrund des Kontextes wird ersichtlich, dass es eigentlich der Flügel des (in der Küche knienden) Engels ist. Durch den bestimmten Artikel wird die Verbindung zu dem bereits erwähnten Engel aufgebaut. Ebenso ist *Auferstehungsengel*[35] durch die Verwendung

des bestimmten Artikels mit dem hier beobachteten ‚Engel' verbunden. Der Beleg *Riemenschneiderengel*[29] steht hier in der Akkusativform im Plural.

Eine andere Art, wie die Sätze verknüpft sind, ist die Beziehung zwischen den Substantiven und Pronomen. Die Pronomen nehmen eine anaphorische Funktion ein.

- Repetition durch Pronomen

Wenn der folgende Textabschnitt als zwei grammatische Sätze interpretiert wird, dann kann hier von einer Satzverknüpfung durch die Repetition der Pronomina (*er*[15]) gesprochen werden: *Der kniende, aus Lindenholz geschnitzte Engel*[14] *blies eine Posaune: so konnte* ***er***[15] *dem Jüngsten Gericht zugeordnet werden*. [...] *Selbst nach dem komplizierten Auftragen der acht leimgebundenen Kreideschichten war* ***ihm***[23]*, wenngleich die Vergolderin keine Feinheiten der spätgotischen Faltenwürfe zuschlämmte, viel von* ***seiner***[24] *möglichen Schönheit vergangen.*

Diese Topikart ist besser sichtbar bei den Belegen *ihm*[23] und *seiner*[24], die in einem Satz stehen. Das Wort ‚Engel' steht nicht direkt im gleichen Satz, und es ist nur aus dem Kontext abzuleiten, dass diese Pronomen auf den Engel verweisen.

In dem Satz: […] *entstand* ***die Figur***[26] *in* ***ihrer***[27] *von anonymer Hand gewollten Schönheit aufs neue.*' verweist das Possessivpronomen *ihrer*[27] auf *die Figur*[26], und so wird der Bezug auf den Engel hergestellt. Da es nicht im getrennten Satz steht, erfüllt es nicht die satzverknüpfende Funktion.

In dem hier behandelten Textabschnitt werden die Referenzen auch durch unterschiedliche Lexeme wiederholt. Eine Übersicht gibt die nächste Topikart an.

- Topiks mit lexematischer Variation – bei dieser Art der Wiederaufnahme sind die konkreten Beziehungen aus dem Kontext zu erschließen:

Sie hatte ***die holzgeschnitzte Figur***[18] *aus anderer Werkstatt übernommen.*
Der auf linkem Knie kniende Engel[19] *muß in seinem geflickten Zustand erbärmlich ausgesehen haben:* [er[21] war] ***ein Veteran***[22] *wechselnder Zeitläufte.*
Dennoch hat er ***die wiederbelebte Arbeit***[30] *künstlerisch nicht allzu hoch eingeschätzt: »Mit anderen Figurationen wird der Engel*[31] ***Beiwerk eines Altars***[32] *gewesen sein, als dessen zentrales Motiv ich die Auferstehung vermute.*

Erstaunlich, wie ***das heruntergekommene Stück***[33] *unter Alexandras Händen gewonnen hat.*

Diese Belege beziehen sich semantisch auf den Referenzbereich ‚Engel' bzw. ‚Engelstatue'. Die Vorkommen Nr. 18 und 33 (u. U. auch Nr. 30) sind die einzigen Wiederholungen in den jeweiligen Sätzen, deswegen kann ihre satzverknüpfende Funktion gut beobachtet werden. In den Sätzen mit den Belegen Nr. 22 und 32 sind auch andere Wiederaufnahmen (durch Rekurrenz) präsent, die auf ‚Engel' hinweisen.

Im untersuchten Textabschnitt kommt noch ein Typ der Wiederaufnahme vor, der nur unter bestimmten Bedingungen als Wiederaufnahme anerkannt werden kann:

- elliptische Topiks – diese Art der Wiederaufnahme kommt hier in diesen Fällen vor:

Der Engel[4] *kniete auf dem Küchentisch:* [**er** war[5]] *einen knappen Meter hoch.*

Der auf linkem Knie kniende Engel[19] *muß in seinem geflickten Zustand erbärmlich ausgesehen haben:* [**er** war[21]] ***ein Veteran***[22] *wechselnder Zeitläufte.*

[…] *verspricht sie sich und mir: Wirst sehen,* [**er**[34]] *wird sein wie neugeboren.*

Diese Fälle können nur dann als satzverknüpfend anerkannt werden, wenn die hier zitierten Textabschnitte jeweils als zwei grammatische Sätze interpretiert werden; dann kommt es zur Wiederaufnahme mit dem Pronomen ‚*er*'.

Die behandelten Belege deuten durch ihre Semrekurrenz und Referenzidentität auf das Referenzobjekt ‚Engel' hin und bilden eine Isotopiekette, die das jeweilige Teilthema erschließt, das sich an dem Textthema beteiligt.

Im Textverlauf haben die Beziehungen zwischen den Wörtern auch eine andere Funktion, als nur die Sätze formal und inhaltlich zu verknüpfen, sie sind nämlich auch im Bereich der Teilthemaentwicklung tätig.

d) Die Untersuchung des ausgewählten Referenzbereiches wird mit der Beobachtung der Themahinweise abgeschlossen. In dieser Phase wird nicht das Gesamttextthema untersucht, sondern ein thematischer Strang, ein Teilthema, dass eine Nebenhandlung in diesem Roman betrifft. In diesem Textabschnitt wird nach Thema-

hinweisen gesucht und jede Referenz auf ‚Engel' selbstständig beobachtet. Es wird von den bisher behandelten Belegen zum Teilthema ‚Engel' ausgegangen:

- Themaeinführungshinweis – das erste Vorkommen des restaurierten Engels, das sich auf dieses Teilthema im Text bezieht. Die Verwendung des unbestimmten (kataphorischen) Artikels dient als ein Signal der Themaeinführung: *einem spätgotisch knienden Engel*[1]
- Themabeibehaltung – *Mit dem Engel*[2] – der bestimmte, der sog. anaphorische Artikel[96] und die Rekurrenz des gleichen Lexems zeigen an, dass es sich nicht um das erstmalige Auftreten des Lexems handelt. Dasselbe auch bei: *Dort kniete spätgotisch der Engel*[3]; *Der Engel*[4]
- elliptische Themabeibehaltung – das einzusetzende Pronomen referiert auf den Engel: [*er*[5] *war*] *einen knappen Meter hoch*
- Themabeibehaltung durch Rekurrenz (auch innerhalb der Wortverbindungen): *das Blattgold auf dem Engelsflügel*[6] *heimwärts in Alexandras Küche zum knienden Engel*[7]; *Du weißt, Engel*[8] *wartet auf uns in Küche*; *Der meterhoch kniende Engel*[9]; *auf dem Altholz des Engels*[10]; *ein spätgotisch kniender Engel*[11]; *den Engel*[12]
- semantischer Themaentwicklungshinweis: *eine anonyme, wahrscheinlich süddeutsche, wenn nicht böhmische Arbeit*[13]
- Themabeibehaltung durch Rekurrenz: *Der kniende, aus Lindenholz geschnitzte Engel*[14]
- Themabeibehaltung durch Pronominalisierung: *so konnte er*[15] *dem Jüngsten Gericht zugeordnet werden.*
- Themabeibehaltung durch Rekurrenz: *ein Chor Engel*[16]; *Alexandras Engel*[17]
- Themaentwicklung durch Substitution: *die holzgeschnitzte Figur*[18]
- Themabeibehaltung durch Rekurrenz: *Der auf linkem Knie kniende Engel*[19]
- Themabeibehaltung durch Pronominalisierung: *seinem*[20]

96 Vgl. Hausendorf / Kesselheim 2008, 115ff.

- elliptische Themabeibehaltung – das einzusetzende Pronomen referiert auf den Engel: [er[21] war] *ein Veteran*

- Themaentwicklungshinweis durch Metapher: *ein Veteran*[22] *wechselnder Zeitläufte* – der Zusammenhang ist nur auf dem Niveau der Bedeutung erschließbar, sowohl der hier behandelte Engel als auch ein Veteran haben das Merkmal ‚alt' bzw. ‚altgedient'

- Themabeibehaltung durch Pronominalisierung: *ihm*[23]; *viel von seiner*[24]

- Themabeibehaltung durch Rekurrenz: *der kniende Engel*[25]

- Themaentwicklung durch Substitution: *die Figur*[26]

- Themabeibehaltung durch Pronominalisierung: *in ihrer*[27]

- Themabeibehaltung durch Rekurrenz: *des blasenden, langgelockten, eher einen jungen Mann als eine Jungfrau im Faltenwurf verbergenden Engels*[28]*; Riemenschneiderengel*[29]

- semantischer Themaentwicklungshinweis – *die wiederbelebte Arbeit*[30] – das verbindende Merkmal von *Arbeit* und *Engel* (als holzgeschnitzte Statue) ist ‚ein durch Betätigung entstandenes Werk'

- Themabeibehaltung durch Rekurrenz: *der Engel*[31]

- semantischer Themaentwicklungshinweis – *Beiwerk eines Altars*[32] – der *Engel* und *Beiwerk* sind durch das Merkmal ‚Schmuck' zusammen verbunden.

- semantischer Themaentwicklungshinweis – *das heruntergekommene Stück*[33] – der Zusammenhang zwischen *Stück* und *Engel* besteht in dem Merkmal ‚einzelner Gegenstand'

- elliptische Themabeibehaltung – das einzusetzende Pronomen referiert auf den Engel: [er[34]] *wird sein wie neugeboren*

- Themabeibehaltung durch Rekurrenz: *vom hochglanzvergoldeten Auferstehungsengel*[35]

Außer den genannten Hinweisen tauchen in den Texten auch wissensabhängige Hinweise auf, die zu den semantischen Hinweisen gehören. Diese Themaentwicklungs-

hinweise gehen zum Beispiel von den Konnotationen aus. Auf diese Weise werden weitere Hinweise im Text, die bisher nicht erwähnt wurden, zu der Aufzählung ergänzt.

Das Weltwissen vermittelt das Bild, dass ein Engel als ein schön aussehender junger Mensch mit Flügeln im luftigen Gewand dargestellt wird, und so ergeben sich weitere Hinweise auf den Engel bei den Belegen wie: *des rechten Flügels; mit beiden Flügeln; Zehenspitzen; Faltenwürfe*. Oder durch die Konnotation gewonnene weitere Teile der Statue: *eine Posaune*; *Posaunentrichter* bzw. das Material der Statue: *Altholz; Lindenholz*.

Es hängt mit dem Weltwissen zusammen, dass alte Statuen nach einer bestimmten Zeit restauriert werden, deswegen könnten in diesem Bereich solche Hinweise aufgelistet werden, die mit der Restaurierung und Vergoldung zusammenhängen. Diese Isotopiehinweise unterstützen den Handlungsrahmen der Vergoldung, zu dem die Bearbeitung der Oberfläche der Statue oder Vergolder-Werkzeuge als Subthemen gehören. Dieser Bereich gehört nicht mehr zum Interesse dieser Analyse.

5.3.2 Abschließende Bemerkungen

Bei der Untersuchung des Lexems ‚Engel' wurden die Schritte der Vorgehensweise bei der Analyse der Teilthemaentwicklung im Gesamttext herausgearbeitet.

Zuerst wurden die Beziehungen zwischen den Wörtern auf der Textoberfläche beobachtet, um die eventuell relevanten Belege zu erfassen. Der Textabschnitt wurde nach den reinen Wiederholungen des gewählten Lexems ‚Engel' durchsucht. Neben den alleine stehenden Belegen kommt das Wort auch in Wortbildungen vor, deswegen ist bei der Suche die Großschreibung der Zeichenkette ‚engel' nicht festgesetzt. Durch die Suche nach den wiederholt auftretenden Belegen wird (mindestens grob) die Grenze des zu untersuchenden Textbereichs gesetzt.

Die Anzahl der Belege wird durch die Pronomen erweitert, die den Engel ersetzen. Und um auch weitere Elemente des Referenzbereichs ‚Engel' sichtbar zu machen, wird nach Zusammenhängen gesucht, die auf gemeinsamen Bedeutungsmerkmalen basieren.

Das Sammeln der Belege aufgrund der unterschiedlichen Wortschatzgruppierungsarten ermöglicht es, die textkonstituierenden Kohäsions- und Kohärenzbeziehungen zwischen den Wörtern zu erkennen. Auch in dieser Phase wird bei der

Suche nach den zusammenhängenden Ausdrücken zuerst von der Textoberfläche ausgegangen.

Die Kohärenzmittel, also das wiederholte Vorkommen der Seme, besonders die von Kontext abhängende Referenzidentität, bilden eine Isotopiekette, die die Beziehungen zwischen unterschiedlichen Wörtern im Text zeigt. Der Kontext ist entscheidend, die Beziehungen zu anderen Wörtern innerhalb des jeweiligen Referenzbereichs anzuzeigen, die auf den ersten Blick nicht zusammenhängen, wie z. B. *Engel – Veteran – Arbeit*.

Die gewonnenen Belege werden in einer Übersicht, die die Anzahl ihrer konkreten Vorkommen angibt, zusammengefasst. Insgesamt wurden in der oben beschriebenen Analyse 43 Themahinweise behandelt (acht davon sind in der Aufzählung nicht nummeriert). Das erste Vorkommen von *Engel* stellt den Themaeinführungshinweis dar. Die Themabeibehaltung wird durch 27 Belege signalisiert (wie z. B. 19 Mal davon als Rekurrenz von *Engel*). Weitere 7 Hinweise (z. B. *Arbeit*, *Veteran*) gehören zur Themaentwicklung. Die restlichen erwähnten 8 Belege (*des rechten Flügels*; *mit beiden Flügeln*; *Zehenspitzen*; *Faltenwürfe*; *eine Posaune*; *Posaunentrichter*; *Altholz*; *Lindenholz*) deuten mehr oder weniger indirekt auf den Engel hin, innerhalb des Kontextes wird die Zugehörigkeit jedoch sichtbar. Diese Untersuchung hat gezeigt, dass in diesem Textabschnitt die Themabeibehaltungshinweise am häufigsten vertreten sind.

In Abhängigkeit vom Untersuchungszweck können bei den Textwörtern unterschiedliche Funktionen betrachtet werden. Die Wörter in den Sätzen deuten nicht nur auf die Zusammengehörigkeit der Sätze innerhalb eines Textes hin, sie zeigen auch das Textthema an. Dabei wird unterschieden, ob das Thema eingeführt, beibehalten, entwickelt oder abgeschlossen wird. Die Teilthemaprogression wird nun im Gesamttext beobachtet.

5.4 Teilthemaentwicklung im Gesamttext

Bei dieser Analyse wird von den Schritten ausgegangen, die bei der Untersuchung des Referenzbereichs ‚Engel' entworfen wurden. Zuerst wird der behandelte Romantext inhaltlich vorgestellt.

5.4.1 Inhalt des untersuchten Textes

Die Kenntnis des Textinhalts ist behilflich, die Beziehungen zwischen den Textwörtern zu erkennen, die oft mehr oder weniger streng an den konkreten Kontext gebunden sind. Deswegen wird auf den Inhalt des untersuchten Romans ‚Unkenrufe' von Günter Grass näher eingegangen.

Diese Beschreibung bezweckt nicht jedes Detail der Handlung zu erfassen, sondern den Zusammenhang zwischen den Textteilen näher zu bringen. Die sieben Kapitel können folgendermaßen zusammengefasst werden:

1. Kapitel: Die Protagonisten, der Witwer Reschke, ein deutscher Kunsthistoriker, und die Witwe Piątkowska, eine polnische Restauratorin, treffen sich auf Allerseelen 1989 zufällig in Gdańsk auf einem Markt, als die Witwe Blumen für das Grab ihrer Eltern kauft. Nach dem Kennenlernen besucht das Paar einen Friedhof, und während ihres Gesprächs kommen sie auf die Idee, die Völker (Polen und Deutschland bzw. auch Litauen) nach dem Zweiten Weltkrieg auf eine eigenartige Weise zu versöhnen.

2. Kapitel: Reschke lernt einen Bengalen, Chatterjee kennen. Chatterjee ist immer positiv gestimmt und spricht über seine Pläne von einem Rikschabetrieb. Reschke und Piątkowska suchen nach einem Grundstück für die Gründung ihres Versöhnungsfriedhofs.

3. Kapitel: Überwiegend in Briefen werden neben persönlichen Angelegenheiten auch Einzelheiten für die Gründung der Gesellschaft besprochen. Es kommen weitere Gesellschafter hinzu, der Aufsichtsrat wird gebildet und die Friedhofsgesellschaft gegründet – zuerst nur als deutsch-polnische Friedhofsgesellschaft, die litauische Komponente soll später hinzutreten.

4. Kapitel: Der Versöhnungsfriedhof wird eingesegnet, und erste Beerdigungen finden statt. Piątkowska sagt zum ersten Mal, dass sie mit dem Projekt aufhören sollen, solange die Gesellschaft noch prosperiert. Es kommen erste Probleme auf, wie z. B. Diebstähle der Blumen und Kränze auf dem Friedhof. Es wird versucht, die Diebe mit

einem Zaun an den Diebstählen zu hindern. Der Zaun ruft negative Reaktionen in der Öffentlichkeit hervor.

5. Kapitel: Es entstehen neue Vorschläge, wie die Friedhofsgesellschaft ihr Angebot erweitern könnte. Einerseits werden Stellen in Seniorenheimen angeboten, damit die Interessenten ihren Lebensabend schon in ihrer Heimat verbringen und auch in ihrer Heimat auf dem Versöhnungsfriedhof begraben werden könnten. Zu dieser Zeit kommt der Gedanke von Umbettungen auf. Als Reschke und Piątkowska zusammen ihre Kinder besuchten, werden sie nicht mit Verständnis für ihre Friedhofsgesellschaft empfangen.

6. Kapitel: Die Gesellschaft hat Erfolg, es gibt ausreichende Nachfrage nach den angebotenen Dienstleistungen. Ein Mitglied des Aufsichtsrates, Frau Erna Brakup, tritt als erste aus dem Aufsichtsrat aus, weil sie nicht mehr damit einverstanden ist, dass sich die Gesellschaft immer mehr auf Geschäfte orientiert. Es wird ein neues Projekt vorgestellt, das Ferienhäuser und Golfplätze für die Familienmitglieder der Verstorbenen an den Orten ihrer Vorfahren anbieten soll. Ein weiteres Mitglied, Jerzy Wróbel, der mit der Landnahme für das neue Projekt nicht einverstanden ist, tritt ab. Der Aufsichtsrat setzt als ein weiteres Projekt durch, die Straßen- und Denkmalschilder mehrsprachig zu gestalten. Dies geschieht aber nicht im Sinne des Versöhnungswerks, sondern mit dem Ziel von Gewinnstreben und Nutzen für Touristen. Reschke und Piątkowska wollen zurücktreten, bleiben aber schließlich doch noch als Ehrenvorsitzende ohne Einspruch- oder Vetorecht. Neue Aufsichtsratmitglieder helfen, die Gesellschaft im geschäftlichen Sinne zu entwickeln.

7. Kapitel: Es werden weitere Versöhnungsfriedhöfe und eine Entbindungs-station mit Kreißsaal für die Trauergäste gegründet. Ein weiteres Mitglied, der Priester Bieroński, tritt zurück. Der Kundendienst erweitert sein Angebot mit neuen Sargformen. Wegen der wachsenden Gewinnsucht der neuen Aufsichtsratmitglieder treten Reschke und Piątkowska zurück, weil die Gesellschaft nicht mehr der ursprünglichen Idee, der Heimkehr der Toten und dem Versöhnungsgedanken, entspricht. Reschke und Piątkowska heiraten, und nach der Feier kommt der Gedanke auf, die Chronik der

Friedhofsgesellschaft zu schreiben. Das Paar tritt seine Hochzeitsreise nach Italien an, doch sie sterben bei einem Autounfall.

Das Ziel der Untersuchung ist nicht das Thema des Gesamttextes zu analysieren, sondern es wird nur ein herausragendes Teilthema betrachtet, das innerhalb des Textthemas zu finden ist.

Aus dem Textinhalt geht hervor, dass sich in diesem Roman eine Liebesgeschichte abspielt. Das Paar lernt sich im ersten Kapitel kennen, die beiden bekommen während eines Friedhofsgespräches eine Idee, wie die Völker nach dem Zweiten Weltkrieg versöhnt werden könnten. Diese Versöhnung wird durch die Friedhofsgesellschaft realisiert. Die Bekanntschaft von Reschke und Piątkowska erreicht ihren Gipfel bei ihrer späteren Hochzeit. Die Liebesgeschichte des Paars ist ein auffälliges Teilthema, das sich eher aus dem Inhalt erkennen lässt, als mit anderen Mitteln, z. B. der Frequenz-Wortformenliste. Die Bestimmung der Teilthemen durch die verwendete Wortformenliste wird in dieser Phase also durch den Textinhalt unterstützt und bestätigt.

Die unten angegebenen Einträge aus der (nicht sortierten) Wortformenliste stellen die ersten zehn Positionen dar, die sich auf die Protagonisten beziehen:

Rank	**Freq**	**Wort**[97]
19	313	Reschke
39	175	Alexandra
63	119	Piątkowska
82	89	Paar
91	83	Witwe
121	60	Reschkes
124	57	Alexander
154	46	Witwer
188	37	Alexandras
231	31	Professor

Die Auswahl dieser Belege sowie der Zusammenhang zwischen ihnen ist streng an den Text des Romans gebunden. Die Beziehungen bei den Eigennamen werden aufgrund der aufgeführten Arten der Wortschatzgruppierung nicht erkannt. Die Aufmerksamkeit richtet sich auf den Referenzbereich ‚Friedhofsgesellschaft' als das zu analysierende Teilthema.

97 Die Zahlen unter ‚Rank' geben die Position in der vollständigen Wortformenliste an, die Angaben bei ‚Freq' geben die Vorkommensfrequenz an (im Gesamttext), und unter ‚Word' stehen die jeweiligen Wortformen.

5.4.2 Referenzbereich und Teilthemaentwicklung von ‚Friedhofsgesellschaft'

Das Teilthema wird für die Untersuchung unter Verwendung der Frequenz-Wortformenliste identifiziert. Die Wortformenliste wurde bereits in einer früheren Phase dieser Arbeit erzeugt und sortiert[98] mit dem Umstand, dass z. B. die Funktionswörter (Artikelwörter, Präpositionen usw.) oder die Namen der Protagonisten ausgelassen wurden[99]. Es hat sich herausgestellt, dass die drei am häufigsten vorkommenden Types[100] in der sortierten Liste sind: *Idee* (72 Belege), *Friedhofsgesellschaft* (63 Belege) und *Friedhof* (43 Belege). Die Frequenz-Wortformenliste und die Kenntnis des Romaninhaltes ergänzen sich gegenseitig bei der Bestimmung der zu untersuchenden Referenzbereiche.

Das am häufigsten vertretene Lexem dieser sortierten Liste ist ‚Idee'. Aus dem Textinhalt geht hervor, dass ‚die Idee' in manchen Fällen auf den Bereich der ‚Friedhofsgesellschaft' referiert. Es enthält die Information, dass die Friedhofsgesellschaft gegründet wird, die der Völkerversöhnung dienen soll. Im bearbeiteten Text sind auch andere ‚Ideen' zu finden. Nicht alle diese Belege deuten also auf die Friedhofsgesellschaft hin; deswegen geht die Untersuchung von der nächsten Stelle in der Wortformenliste aus, von *Friedhofsgesellschaft*, weil dieser Eintrag eindeutig ist.

Wenn innerhalb des Dekompositums ‚Friedhofsgesellschaft' das Bestimmungswort *Friedhof* und das Grundwort *Gesellschaft* getrennt betrachtet werden, dann tragen sie beide wesentlich dazu bei, die Beziehungen der anderen Textwörter zu dem Referenzbereich *Friedhofsgesellschaft* zu entdecken.

Weitere Elemente, die zu diesen Referenzbereichen gehören, können nach den Prinzipien ausgewählter Wortschatzgruppierungsarten gefunden werden, diese Gruppierungen sind in dieser Arbeit die Wortfamilien, Sachgruppen und Wortfelder. Die Betrachtung des Wortschatzes nach diesen Regeln hilft, Kohärenzbeziehungen in den festgelegten Referenzbereichen zu identifizieren. Diese Beziehungen deuten nicht nur auf Isotopien im Text hin, sondern zeigen auch das Thema bzw. das Teilthema im Text an. Die Beschreibung der Kohäsions- und Kohärenzmittel dieses Teilthemas findet in diesem Abschnitt nicht mehr statt, sondern es werden nur die Prinzipien dieser

98 Vgl. das Unterkapitel „4.1.1 Wortformenliste".

99 Die Wortformenliste ist auf der beigelegten CD in der Datei ‚`grass_de_wortliste_auswahl_ohne_doppelformen.txt`' zugänglich.

100 Unter einem ‚Type' wird zusammenfassend das Auftreten mehrerer Token verstanden.

Beschreibung dazu verwendet, die Zusammenhänge zwischen den Wörtern im Text ausfindig zu machen.

Es ist nicht mehr nötig, diese Wortschatzbeziehungen einzeln für die Schlüsselwörter ‚Friedhofsgesellschaft', ‚Friedhof' und ‚Gesellschaft' zu behandeln, das Prinzip wurde in den bereits durchgeführten Untersuchungen angedeutet (z. B. für das Wort ‚Friedhof' im Unterkapitel „4.2.3 Wortbildung und Text"). In dieser Phase wird nicht von einer Auflistung der Wörter, die z. B. zu einer Wortfamilie gehören, ausgegangen. Im Zentrum der Betrachtungen stehen Wörter, so wie sie im Text gebraucht werden und ihre konkreten Beziehungen.

Für die Untersuchung werden die Instrumente der Themabeschreibung verwendet (so wie sie im Unterkapitel „5.2 Themahinweise" vorgestellt wurden), jedoch mit dem Unterschied, dass nicht das Gesamttextthema im Mittelpunkt der Betrachtung steht, sondern nur eine Untermenge, ein Teilthema. Die einzelnen Hinweise auf die Phasen des Auftretens eines Teilthemas im Text werden in der Rekapitulation der Themahinweise zusammengefasst. Die Untersuchung konzentriert sich auf folgende Hinweise:

1) Themaeinführungshinweise – der unbestimmte (kataphorische) Artikel. Bei den Teilthemen ist die Einführung auch das erste Vorkommen der Referenz, die sich auf den jeweiligen Referenzbereich bezieht.

2) Themabeibehaltungshinweise realisieren sich durch den bestimmten (anaphorischen) Artikel, die Pronominalisierung, die elliptischen Hinweise und besonders durch die Rekurrenz derselben Lexeme.

3) Themaentwicklungshinweise beziehen sich auf eine vorher eingeführte Referenz, die sie fortführen, differenzieren und ausbauen. Dies geschieht mit den Mitteln der Substitution (Ko-Referenz), mit der semantischen Beziehung anhand des lexikalischen Zusammenhangs (der von dem konkreten Text ausgeht) und mit Referenzen aufgrund der semantischen Beziehung (Wiederkehr semantischer Merkmale als Isotopiehinweise oder als Situations- und Handlungsrahmen).

4) Themaabschlusshinweise werden entweder durch metakommunikative Hinweise, oder durch die Absenz der Themahinweise im weiteren Text realisiert.

Die Aufmerksamkeit konzentriert sich in diesem Abschnitt auf die thematische Progression bzw. auf die Entwicklung des Teilthemas. Im Vordergrund der Betrachtung des Textes stehen die textbezogene Aktivierung semantischer Merkmale, die Sem-rekurrenz und die Referenzidentität. Das Ziel ist, vor allem die Themahinweise, besonders die Teilthemaentwicklung des ausgewählten Referenzbereiches, zu beobachten. Die lexikalischen Beziehungen sind oft nicht ausreichend, diese Beziehung zu entdecken, und die konkrete Verwendung im jeweiligen Text muss bewertet werden.

Bei der Untersuchung werden nur solche Hinweise berücksichtigt, die direkt mit dem untersuchten Referenzbereich, also mit dem Teilthema ‚Friedhofsgesellschaft' verbunden sind. Das bedeutet, dass es sich bei einem ‚Friedhof' um den Friedhof handeln kann, den die Friedhofsgesellschaft betreibt. Im Text kommen nämlich auch andere Friedhöfe vor, die aber zu eigenen Subthemen gehören und mit dem untersuchten Teilthema keinen Zusammenhang haben. Jedes Vorkommen des Lexems ‚Friedhof' kann als Kohäsionsmittel im Text dienen oder zu dem Thema des Gesamttextes beitragen. Die Zusammenhänge werden jeweils durch den Kontext bestätigt oder widerlegt.

Ein Beleg wie ‚Grab' wird nicht mehr zum untersuchten Referenbereich gezählt, weil er auf ‚Friedhof' referiert und zum Subthema ‚Friedhof' gehört. Eine eventuelle Referenz auf Friedhofsgesellschaft ist also über ‚Friedhof' vermittelt.

Die Handlung des untersuchten Romantextes wird an mehreren Stellen der Untersuchung erwähnt, damit die Bedeutung der jeweiligen Belege im konkreten Textgebrauch sichtbar wird. Die Untersuchung wird nach den Romankapiteln gegliedert. Es wird nicht der Text in voller Länge analysiert, sondern ausgewählte Kapitel, um die besprochenen Themahinweise im Text stellvertretend, aber ausreichend zu beschreiben. Es werden vor allem Themaeinführung, -beibehaltung und -abschluss untersucht.

Das Lexem ‚Friedhofsgesellschaft' kommt zum ersten Mal auf der Seite 40 vor, aber schon an früheren Textstellen werden Referenzen auf diesen Bereich gefunden. Dem Textinhalt wurde entnommen, dass sich das Lexem *Idee* auf die Friedhofs-gesellschaft bezieht. Im Text können auch andere Vorkommen von *Idee* beobachtet werden; sobald sie nicht auf ‚Friedhofsgesellschaft' referieren, werden sie von der

Untersuchung ausgelassen. Alle Belege von *Idee*, die auf das Teilthema ‚Friedhofsgesellschaft' referieren, gehören zu einem Subthema.

Zuerst kommt *Idee* im Tagebuch von Reschke vor: […] *»vom Eigentlichen ablenken, von der **Idee**, von unserer großen, die Völker versöhnenden **Idee** ...«*. (GGU, S. 16) Dies ist der erste Beleg des Lexems *Idee* in diesem Text, und durch das erstmalige Vorkommen wird es als Themaeinführungshinweis bezeichnet. Dieser Textabschnitt ist ein Teil eines Satzes im Tagebuch von Reschke, das ein Freund – der Erzähler – mit anderen Unterlagen retrospektiv als Chronik der Friedhofsgesellschaft bearbeitet. Aus der Perspektive des Erzählers ist die *Idee* schon bekannt (Verwendung des bestimmten Artikels), obwohl die Handlung des Romans gerade beginnt. Durch das Auslassungszeichen wird beim Leser ein Erwartungssignal geweckt, und das Teilthema bzw. der Referenzbereich wird eingeführt. Das zweite Vorkommen von *Idee* ist ein Hinweis auf die Entwicklung des Teilthemas durch Substitution.

Der nächste Beleg von *Idee* ist in einer Szene am Anfang des ersten Kapitels zu finden, als sich die beiden Protagonisten kennenlernen und als sie gemeinsam den Friedhof besuchen, auf dem die Eltern von Piątkowska begraben sind. Sie reden davon, dass viele Menschen aus unterschiedlichen Gründen nicht in ihrer Heimat begraben wurden. Die Idee der Gründung der Friedhofsgesellschaft wird wieder erwähnt: *Vielleicht nahm ihre **Idee** erste Gestalt an, um sich mit dem Zigarettenrauch wieder zu verflüchtigen.* (GGU, S. 29) In diesem Satz wird *Idee* als Entwicklungshinweis des Teilthemas durch Substitution interpretiert.

Ein ähnlicher Fall ist in der weiteren Handlung zu beobachten. Reschke und Piątkowska bereiten das Essen in der Wohnung von Piątkowska zu, und es wird weiter von der Situation gesprochen, dass viele Menschen in der Vergangenheit nicht in ihrer Heimat begraben werden konnten: *Und jetzt erst klickte es, fiel der Groschen, wurde, ohne Schmerz, ein **Gedanke** geboren, gelang es dem Witwer und der Witwe, eine **Idee** abzustimmen,* [...] (GGU, S. 37) Die Belege *Gedanke* und auch *Idee* werden als Entwicklungshinweis des Teilthemas durch Substitution interpretiert.

Die Elemente auf dem Niveau eines Subthemas können auf zweierlei Weisen interpretiert werden. Im Rahmen des jeweiligen Subthemas würde die *Idee* bei ihrem wiederholten Vorkommen als Themabeibehaltung durch Rekurrenz bezeichnet, und *Gedanke* könnte als Themaentwicklung durch Substitution eingeordnet werden – es ist aber nicht das Ziel dieser Untersuchung, einzelne Subthemen zu analysieren. Für diese Untersuchung werden nur die Hinweise auf das übergeordnete Teilthema betrachtet; wenn *Idee* oder *Gedanke* auf die *Friedhofsgesellschaft* referieren, handelt es sich deswegen um Themaentwicklung durch Substitution.

Der nächste Satz befindet sich in der unmittelbaren Nachbarschaft: *Es muß ein langes Gespräch bei immer neu aufgebrühtem Kaffee gewesen sein, das* [...] *diese* ***Idee*** *entfacht, zu ihrer* ***Idee*** *erklärt und schließlich zur völkerversöhnenden* ***Idee*** *gewölbt hat*. (GGU, S. 37) Alle drei Belege von *Idee* referieren auf die Friedhofsgesellschaft als Entwicklung des Teilthemas durch Substitution.

Jetzt, nachdem die ***Idee*** *raus ist, kann ich nicht mehr zurück*. (GGU, S. 39) und *Das war später, als ihre* ***Idee*** *schon wie selbsttätig um sich griff*. (GGU, S. 40) *Zwar ging es, kaum war die* ***Idee*** *flügge, nur noch um* ***Begräbnisinstitute****, fristgerechte* ***Bestattungen****,* ***Leichentransport*** *und zu erwartende Schwierigkeiten beim* ***Überführen der Särge und Urnen****, doch fand Alexandra all das und auch den Namen ihrer gemeinsamen* ***Idee****, den Alexander vorgeschlagen hatte, ein Gelächter, ihr Glockenvogelgelächter wert.* (GGU, S. 40) – in allen Fällen des Vorkommens von *Idee* liegt eine Entwicklung des Teilthemas durch Substitution vor.

Die Belege *Begräbnisinstitute*, als ein Unternehmen auf ähnlichem Gebiet, und *Bestattungen*; *Leichentransport, Überführen der Särge und Urnen* als die möglichen angebotenen Dienste der Friedhofsgesellschaft referieren auf den Bereich ‚Friedhofsgesellschaft'[101] als semantische Themaentwicklung.

Im Text, der auf die oben beschriebenen Belege von *Idee* unmittelbar folgt, wird direkt zu den Belegen des Referenzbereichs FG gewechselt: *Sein Vorschlag hieß: »****Polnisch-Deutsche Friedhofsgesellschaft****.«* (GGU, S. 40) In diesem Satz steht das erste Vorkommen von *Friedhofsgesellschaft* im Text in dieser Form als ein Hinweis, der

101 Weiter in der Beschreibung des Referenzbereiches ‚Friedhofsgesellschaft' als ‚FG' abgekürzt.

auf den untersuchten Referenzbereich ‚Friedhofsgesellschaft' referiert. Streng genommen steht an dieser Stelle ein anderes Lexem als bei den vorherigen Fällen, und so könnte dieser Fall als Themaentwicklung durch Substitution betrachtet werden. Der Anfang der Isotopie befindet sich bei dem ersten Vorkommen von *Idee* im Sinne von ‚die Idee, die Friedhofsgesellschaft zu gründen' auf der Seite 16.

Weitere Vorkommen von FG finden sich gleich im unmittelbar folgenden Text: bei *Deutsch-Polnische Friedhofsgesellschaft* (GGU, S. 41) und *die Polnisch-Deutsch-Litauische Friedhofsgesellschaft* (GGU, S. 41) kommt es in beiden Fällen zur Wiederholung des Lexems *Friedhofsgesellschaft,* und es findet eine Themabeibehaltung durch Rekurrenz statt.

Bei dem Kurzwort *PDLFG* (GGU, S. 41) ist in seiner Vollform auch die Wiederholung der Referenz *Friedhofsgesellschaft* vertreten, dieser Fall wird ebenso als Themabeibehaltung durch Rekurrenz interpretiert.

In diesem Text macht sich Reschke Gedanken darüber, was alles bei der Gründung der Gesellschaft nötig ist. Durch das Weltwissen ist es zu erschließen, welche anderen Erfordernisse oder welcher Zubehör zur Gründung einer Gesellschaft gehören, einige davon sind gleich im Text als Entwicklungshinweise des Teilthemas ‚Friedhofs-gesellschaft' vertreten: […] *noch fehlten als notwendiges Zubehör weitere* ***Gründungs-mitglieder****, ein* ***Gesellschaftervertrag****, die* ***Satzung*** *und* ***Geschäftsordnung****, der* ***Aufsichtsrat*** *und – weil auf dieser Welt nichts umsonst ist – das* ***Gründungskapital*** *samt* ***Kontonummer***. (GGU, S. 41) Diese unterschiedlichen Referenzausdrücke werden als semantische Entwicklungsinweise des Teilthemas interpretiert, weil sie inhaltlich zu einem Referenzbereich gehören.

Das erste Kapitel wird damit abgeschlossen, dass die Idee, der Gedanke der Friedhofsgesellschaft entstanden ist und der Entschluss, sie zu gründen, gefasst wurde. Im zweiten Kapitel wird dieses Bestreben weiter entwickelt, besonders in dem Sinne, dass das Paar ein Gelände für den zu gründenden Versöhnungsfriedhof sucht.

Am Anfang des zweiten Kapitels trifft sich Reschke zufällig mit Herrn Chatterjee, der ein Rikscha-Verkehrsunternehmen plant, und beide unterhalten sich über

die Verkehrslage in Europa. Später spricht der Erzähler davon, wie Reschke den verlaufenen Tag in seinem Tagebuch beschreibt:

Erst nachdem er umständlich den ***Friedhof*** *am Hagelsberg beschrieben hat, kommt er zur Sache, wird ihm die von der Witwe vorgeschlagene »****Polnisch-Deutsch-Litauische Friedhofsgesellschaft****« wichtig. Er nennt* ***sie*** *»unsere, kaum gezeugt, schon geborene* ***Idee****«.* ***Die litauische Komponente*** *wertet er als »einleuchtend und wünschenswert«, zugleich aber als »schwer in die Tat umzusetzen«. Und doch gibt er der Schubkraft* ***des Unternehmens*** *Vorschuß: »****Das Projekt Wilna*** *wird zu finanzieren sein.* (GGU, S. 53)

Der Beleg *Friedhof* wird hier angeführt, um zu illustrieren, dass nicht alle Belege, die zu dem Bereich ‚Friedhof' gehören (z. B. im Rahmen der gleichen Wortfamilie) auch auf die FG hinweisen. Dieser Beleg deutet auf den vorher besuchten Friedhof hin, es zeigt nicht auf den Referenzbereich der FG. Weitere hervorgehobene Belege in diesem Abschnitt referieren auf die FG als Beibehaltung des Teilthemas durch Rekurrenz: *Polnisch-Deutsch-Litauische Friedhofsgesellschaft* und durch pronominale Beibehaltung *sie*. Gleich darauf folgt auch eine Entwicklung durch Substitution mit *Idee*. Die Substitution der FG durch den Ausdruck *des Unternehmens* sorgt auch für die Teilthemaentwicklung. Die gebliebenen Belege *Die litauische Komponente* und *Das Projekt Wilna* werden beide als die semantische Entwicklung bewertet. Sie beziehen sich inhaltlich als Teile dieses Projektes auf die Friedhofsgesellschaft.

»Hör zu, Reschke«, hab' ich mit seinem Füller an den Rand gekritzelt, »das ist eine ***Furzidee****!«* (GGU, S. 53) Der Ausdruck *Idee* kommt hier als ein Hinweis auf die FG vor, es handelt sich um Entwicklung durch Substitution. Die erste Konstituente drückt die Bewertung oder Einschätzung der Idee als unsinnig aus.

[...] *die* ***Kosten*** *der* ***Friedhofsgesellschaft***. (GGU, S. 54) Bei dem Lexem *Kosten* wird durch das Weltwissen begründet, dass beim Betrieb einer Gesellschaft auch unterschiedliche Kosten entstehen, es handelt sich also um einen semantischen Entwicklungshinweis. Der Beleg *Friedhofsgesellschaft* stellt Themabeibehaltung durch Rekurrenz dar.

Die Protagonisten treffen sich, um die Gründung der FG näher zu besprechen und nach einem Grundstück für den zu gründenden Friedhof zu suchen:

*Man wollte geeignetes **Gelände** finden und der **Idee** als **Projekt** zu fester Kontur verhelfen,* […] (GGU, S. 62) Der Beleg *Gelände* wird als ein semantischer Themaentwicklungshinweis betrachtet. Durch den Kontext wird nämlich deutlich, dass das Gelände als der zukünftige Friedhof der FG dienen wird. *Idee* und *Projekt* referieren auf die FG und kommen in der Funktion der Themaentwicklung durch Substitution vor.

Während der Suche nach dem Ort für den zu gründenden Friedhof werden mehrere Stellen besichtigt: *Reschke hält fest, daß er das **Gelände** zwischen Brentau und Matern anfangs als geeignet für einen **Waldfriedhof** gesehen habe, zumal »die hochstämmigen Buchen gut Abstand halten und dem **Projekt** einen natürlichen Rahmen leihen.* […] *Gegen diese ideale **Friedhofslage** sprach der zu nah gelegene Flughafen von Gdańsk* […] (GGU, S. 64)

Bei den Ausdrücken *Gelände* und *Friedhofslage* liegt semantische Themaentwicklung vor, bei beiden Belegen ist die Beziehung zur FG durch den Kontext zu begründen. Der *Waldfriedhof* wird als semantische Entwicklung bewertet, denn durch das Weltwissen kann erschlossen werden, dass die FG auch einen Waldfriedhof betreiben kann. Das *Projekt* wird als semantische Entwicklung interpretiert, weil es das Vorhaben der FG vertritt.

[…] *ihre **Idee** lädiert, ihr Entschluß, kaum gefaßt, schon gebremst,* […] (GGU, S. 65) Das Lexem *Idee* ist ein Entwicklungshinweis auf die FG durch Substitution.

*»Ich meine, inzwischen ist diese **Anlage** ...«* (GGU, S. 65) Der Beleg *Anlage* wird als ein semantischer Entwicklungshinweis betrachtet. Der Kontext zeigt, dass diese Anlage als der zukünftige Friedhof der FG verwendet werden soll.

Auf die gleiche Weise werden die anderen Belege von ‚Anlage' oder ‚Gelände' in diesem Textabschnitt (GGU, S. 65–69) bewertet – wenn es durch den Kontext als ein Grundstück für den neuen Friedhof vorgesehen ist, dann werden die Lexeme als semantische Themaentwicklung interpretiert.

Nach der Besichtigung des Geländes für den zukünftigen Friedhof kommt das Paar zu dem Entschluss, wo ihr Versöhnungsfriedhof gegründet werden soll:

*»**Hier wird sein!**« rief Alexandra Piątkowska* [...] (GGU, S. 69) Bei *Hier wird sein!* handelt es sich um eine Ellipse, das Subjekt fehlt und dem Kontext kann entnommen werden, dass hier ‚der Friedhof' einzusetzen ist. Der Kontext zeigt, dass hier der zu gründende Friedhof gemeint ist und dass sich dieser Fall auf das Teilthema ‚Friedhofsgesellschaft' bezieht. Dieser einzusetzende Beleg *Friedhof* wird als semantischer Entwicklungshinweis auf die FG interpretiert.

[…] *Umsetzbarkeit ihrer **Idee** gewiß, sagte sie, als das Paar endlich die **Parkanlage** verließ* […] (GGU, S. 70) Bei dem Lexem *Idee* handelt es sich um ein weiteres Entwicklungssignal durch Substitution. Das Lexem *Parkanlage* referiert auf den zukünftigen Friedhof und ist so als semantische Entwicklung anzusehen.

Im Text wird das Gespräch über die Gründung der FG auch auf dem Rückweg von dem besuchten Gelände weitergeführt:

[…] *dann auf dem Weg zum geparkten Wagen sprach Alexandra Piątkowska vom Erlös ihrer fast schlaflosen Nacht, vom **Geld**, das als **Startkapital** nötig sein werde.* (GGU, S. 70) Die Lexeme *Geld* und *Startkapital* werden als semantische Entwicklungshinweise des Teilthemas bezeichnet, die durch das Weltwissen über die Gründung einer Gesellschaft auf die FG deuten.

*Da der **Złoty** nichts tauge, müsse die **Währung des westdeutschen Staates** ihrer zum **Projekt** gewordenen **Idee** das Fundament legen*. (GGU, S. 70) *Mit **Deutschmark** wird klappen*. (GGU, S. 70)

Durch das Weltwissen kann die Bezeichnung von *Złoty*, *Währung des westdeutschen Staates* und *Deutschmark* als ‚Kapital für die Gründung einer Gesellschaft' als Bezug auf die FG gewertet werden. Diese Belege werden als semantische Entwicklung des Teilthemas interpretiert. Bei den Lexemen *Projekt* und *Idee* handelt es sich um Themaentwicklung durch Substitution von FG.

*Sie sagte: »Eine Million **Deutschmark** zu Anfang muß sein.«* (GGU, S. 70) und […] *Alexandras Hinweis auf ausreichend viel **Deutschmark*** […] (GGU, S. 72) Bei

Deutschmark liegt semantische Themaentwicklung vor, es hat die Bedeutung des Kapitals für die Gründung einer Gesellschaft, und so referiert der Beleg auf die FG.

Es sei ihm, meldet Reschke, die Witwe um den Hals gefallen, als ihr, vom Rondell aus gesehen, der zukünftige ***Friedhof*** *vor Augen stand.* (GGU, S. 71) Innerhalb einer Mehrwortverbindung kann die Bedeutung präzisiert und eindeutiger aufgenommen werden, so dass *der zukünftige Friedhof* zu dem Referenzbereich von FG als semantische Themaentwicklung gerechnet wird.

Jetzt fehlt nur noch ***Platz*** *in Wilno!* (GGU, S. 71) Das Lexem *Platz* stellt ein Entwicklungssignal dar, denn durch den Kontext kann der semantische Zusammenhang mit einem der zukünftigen, von der FG betriebenen Friedhöfe rekonstruiert werden.

Auf jeden Fall verlangt unser ***Doppelprojekt*** *ganzen Einsatz!«* (GGU, S. 72) Das Lexem *Doppelprojekt* referiert direkt auf die FG als Entwicklung durch Substitution.

In der Hotelbar sehen Reschke und Piątkowska eine Touristengruppe aus Deutschland; die Witwe sagt dabei:

Werden bald ***Kunden*** *sein alle*. (GGU, S. 73) Und weiter im Text steht: *Sie wäre womöglich – nach nur zwei Whisky – mit direkter* ***Kundenwerbung*** *aktiv geworden,* [...] (GGU, S. 73) Das Weltwissen zeigt, dass eine Gesellschaft (hier die FG) ihre *Kunden* hat und dass die *Kunden* durch Werbung (*Kundenwerbung*) gewonnen werden. Bei diesen Belegen geht es um semantische Entwicklung.

Sprachen sie zwischendurch über ihre ***Idee****?* (GGU, S. 79) Der Beleg *Idee* ist ein Entwicklungshinweis durch Substitution.

War, zumindest beiläufig, von ***Friedhöfen*** *in Gdańsk und Wilna, von ausreichend viel* ***Deutschmark*** *die Rede?* (GGU, S. 79) In diesem Beispiel werden die *Friedhöfe* mit der Ortsangabe präzisiert, durch den Kontext wird also begründet, dass es sich um die Friedhöfe handelt, die die FG betreiben wird. So wird der Beleg *Friedhöfen* als semantische Entwicklung des Teilthemas interpretiert. Der Beleg *Deutschmark* im Sinne des Gründungskapitals deutet auf die FG als semantische Teilthemaentwicklung hin.

Oder blieb neben der Liebe im schmalen Bett kein Platz für ***Friedhöfe*** *hier und dort?* (GGU, S. 79) Die *Friedhöfe hier und dort* werden inhaltlich wie die vorher erwähnten *Friedhöfe in Gdańsk und Wilna* interpretiert. Weil der Kontextbezug auf die FG realisiert wird, wird dieser Fall ebenso als semantische Entwicklung des Teilthemas interpretiert.

Oder wurde ihre ***Idee*** […] *durch Liebe beatmet?* (GGU, S. 79) Das Lexem *Idee* stellt eine Themaentwicklung durch Substitution dar.

Die Handlung des zweiten Kapitels wird von den Vorbereitungen und der Gründung der Friedhofsgesellschaft geprägt. Im dritten Kapitel wird die Gesellschaft endlich gegründet, bei der Untersuchung wird dieser Abschnitt übergangen, weil bereits ausreichende Hinweise auf Entwicklung und Beibehaltung des untersuchten Teilthemas belegt wurden.

Im vierten Kapitel wird der Versöhnungsfriedhof eingesegnet, und die ersten Beerdigungen finden statt. Der Aufsichtsrat wird tätig, bestimmte Probleme müssen bewältigt werden, aber im Allgemeinen prosperiert die Gesellschaft.

[…] *sie habe sich bei den* ***Verhandlungen*** *durch damenhaftes Lächeln und blitzschnelles Kopfrechnen ausgezeichnet,* […] (GGU, S. 123). Durch das Weltwissen wird der Zusammenhang zu FG deutlich, weil eine Gesellschaft *Verhandlungen* führt. Dieser Beleg wird als semantischer Themaentwicklungshinweis bezeichnet.

[…] *sobald Marczak den* ***Aufsichtsrat*** *rufen wird*. (GGU, S. 123) […] *mit den polnischen* ***Mitgliedern der Aufsicht****,* […] (GGU, S. 123) Die Belege *Aufsichtsrat*, *Mitgliedern der Aufsicht* bzw. *Aufsicht* allein referieren auf FG als organisatorische Einheiten der Gesellschaft und werden als semantische Themaentwicklung interpretiert.

Doch unsere ***Idee****,* […] (GGU, S. 125) Das Lexem *Idee* referiert als Themaentwicklung durch Substitution auf die FG.

Gleichwohl sollten wir bei der gärtnerischen Pflege des ***Friedhofs*** [...] (GGU, S. 125) und […] *geeignet für unsere* ***Friedhofsanlage*** *...«* (GGU, S. 125) Bei den

Belegen *Friedhofs* und *Friedhofsanlage* handelt es sich um den Friedhof, der von der FG betrieben wird. Anhand dieses Zusammenhangs werden diese Belege als semantische Themaentwicklungshinweise interpretiert.

[…] *den Redefluß einer Frau* […] *die als polnische Staatsbürgerin deutscher Herkunft zum* ***Aufsichtsrat*** *der* ***Deutsch-Polnischen Friedhofsgesellschaft*** *gehörte*. (GGU, S. 128) Der *Aufsichtsrat* ist mit der FG durch das Weltwissen verbunden und wird als semantische Themaentwicklung bezeichnet. *Friedhofsgesellschaft* stellt eine Themabeibehaltung durch Rekurrenz dar.

Sie sprach eine aussterbende Sprache, »weshalb ihr« […] *»der* ***Sitz im Aufsichtsrat*** *der* ***Friedhofsgesellschaft*** *zu Recht zugesprochen wurde*. (GGU, S. 129) Der *Sitz* wird durch den Kontext als *Sitz im Aufsichtsrat der Friedhofsgesellschaft* näher bestimmt und als semantische Themaentwicklung interpretiert. Der Beleg *Aufsichtsrat* ist durch das Weltwissen als ein Organ der Friedhofsgesellschaft zu erkennen und wird als semantische Themaentwicklung angesehen. Bei *Friedhofsgesellschaft* handelt es sich um Themabeibehaltung durch Rekurrenz.

Kaum hatten die ***Verhandlungen*** *zur* ***Gründung*** *der* ***Friedhofsgesellschaft*** *begonnen* [...] (GGU, S. 129) Das Lexem *Verhandlungen* wird durch das Weltwissen im Sinne einer allgemeinen Tätigkeit einer Gesellschaft als semantische Themaentwicklung interpretiert. Der Zusammenhang von *Gründung* und FG ist durch den Kontext zu erschließen, dieser Fall wird ebenso als semantische Themaentwicklung betrachtet. Das wiederholte Vorkommen von *Friedhofsgesellschaft* deutet auf Themabeibehaltung durch Rekurrenz hin.

[...] *selbst Reschke hielt sich zurück, obgleich ihm bewußt war, daß schon die Ankündigung des Staatsbesuches dem* ***Unternehmen*** *»****Versöhnungsfriedhof****« ein günstiges Licht geworfen hatte*. (GGU, S. 130) Die Friedhofsgesellschaft wird in diesem Kontext als *Unternehmen* bezeichnet, es handelt sich also um Themaentwicklung durch Substitution mit einem Synonym. Das Lexem *Versöhnungsfriedhof* wird durch den Kontext als der von der FG zu gründende Friedhof verstanden und als semantische Entwicklung interpretiert.

Vielleicht wär' sonst schiefgegangen mit ***Friedhof*** *für Deutsche*. (GGU, S. 130) Der Zusammenhang zwischen dem hier erwähnten *Friedhof* und der FG beruht auf dem Kontext – es liegt semantische Themaentwicklung vor.

Die ***Aufsicht*** *der* ***Friedhofsgesellschaft*** *hatte noch während der* ***Gründungssitzung*** *den geschäftsführenden* ***Gesellschaftern*** *eine pauschale Vergütung zugestanden;* […] (GGU, S. 130)

Die Lexeme *Aufsicht, Gründungssitzung* und *Gesellschaftern* werden als semantische Entwicklung der FG bezeichnet; bei *Friedhofsgesellschaft* liegt Themabeibehaltung durch Rekurrenz vor.

Er richtete in seiner Junggesellenwohnung ein ***Sekretariat*** *ein* [...] (GGU, S. 132) Das Lexem *Sekretariat* stellt eine semantische Themaentwicklung dar.

[…] *denn während der zweiten Junihälfte sollte der* ***Versöhnungsfriedhof*** *feierlich eingesegnet werden: die ersten* ***Begräbnisse*** *standen auf dem Programm.* (GGU, S. 132) Der *Versöhnungsfriedhof* ist durch den Kontext als der Friedhof der FG zu bestimmen und wird als semantische Themaentwicklung interpretiert. Der Kontext bei *Begräbnisse* verweist auf die inhaltliche Beziehung, es sind die ersten Begräbnisse, die die FG durchführt.

[...] *segneten Hochwürden Bieroński* […] *und Konsistorialrat Karau* […] *den* ***Versöhnungsfriedhof*** *als ökumenisches Doppel ein; in Gestalt des katholischen und des evangelischen Geistlichen spiegelte sich nicht nur die konfessionelle Mischung des* ***Aufsichtsrates***, […] (GGU, S. 133)

Die Lexeme *Versöhnungsfriedhof* und *Aufsichtsrates* werden als semantische Themaentwicklung klassifiziert.

[…] *die Einsegnung* ***des Friedhofs*** *und – aus gebotener Distanz – die ersten Beerdigungen aufzuzeichnen*. (GGU, S. 134) Der Kontext zeigt, dass es sich um den von der FG betriebenen *Friedhof* handelt, und deswegen wird dieser Fall als semantische Themaentwicklung interpretiert.

Auf Sommeranfang fanden die Einsegnung und danach zwei ***Beerdigungen*** [...] *statt, dort, wo die Allee zum Hauptgebäude der Technischen Hochschule dem* ***Versöhnungsfriedhof*** *die Grenze zieht*. (GGU, S. 134)

Die *Beerdigungen* werden als Dienstleistungen der FG angeboten und in diesem Sinne als semantische Themaentwicklung angesehen. Bei *Versöhnungsfriedhof* liegt semantische Themaentwicklung vor.

[…] *Fügung, bei der Terminierung der* ***Erstbegräbnisse*** *im Spiel gewesen sei* […] (GGU, S. 135) Die FG bietet Begräbnisse als Dienstleistungen an, und deswegen wird das Lexem *Erstbegräbnisse* als semantische Themaentwicklung klassifiziert.

[…] *zur Stunde der Einsegnung des* ***Versöhnungsfriedhofes*** (GGU, S. 135) Der Beleg *Versöhnungsfriedhofes* drückt eine semantische Themaentwicklung aus.

Die Lage des ***Friedhofs*** *sollte in Augenschein genommen* […] *werden.* (GGU, S. 136) Der Kontext beweist, dass *des Friedhofs* sich hier auf den von der FG betriebenen Friedhof bezieht, und dieser Fall wird als semantische Themaentwicklung bewertet.

Reschke ist noch in seinen Aufzeichnungen bewegt von den ersten ***Bestattungen***. (GGU, S. 136) Das Weltwissen zeigt, dass *Bestattungen* von der FG organisiert werden, deswegen wird das Lexem als semantische Themaentwicklung angesehen.

[…] *die ersten* ***Beerdigungen*** *auf dem* ***Versöhnungsfriedhof*** […] und weiterer Beleg bei: […] *die von den* ***Beerdigungen*** *kaum Notiz nehmen.* (GGU, S. 137) Bei den Lexemen *Beerdigungen* und *Versöhnungsfriedhof* liegt eine semantische Themaentwicklung vor, die auf dem inhaltlichen Zusammenhang beruht.

[…] *die zukünftige Nutzung der* ***Parkanlage*** *als »****Versöhnungsfriedhof****« – »****Cmentarz Pojednania****« – bekannt macht.* (GGU, S. 137) Bei den Belegen *Parkanlage*, *Versöhnungsfriedhof* und *Cmentarz Pojednania* wird durch den Kontext angezeigt, dass es sich jeweils um den von der FG betriebenen Friedhof handelt, und sie werden als semantische Themaentwicklung klassifiziert. Derselbe Fall liegt auch in diesem Belegsatz vor: […] *den Hinweg vom Hotel zum* ***Friedhof*** […] (GGU, S. 137)

Abä ***Beärdigung*** *häd miä jefallen,* […][102] (GGU, S. 138) Bei ‚Beerdigung' (*Beärdigung*) wiederholt sich semantische Themaentwicklung.

[...] *es könne der* ***Versöhnungsfriedhof*** *eng, eines Tages voll, überfüllt sein?* (GGU, S. 139) Das Lexem *Versöhnungsfriedhof* wird als semantische Themaentwicklung interpretiert.

[…] *weiteren* ***Begräbnissen*** *das Geleit gegeben* […] (GGU, S. 139) Das Lexem *Begräbnissen* stellt eine semantische Themaentwicklung dar.

[…] *die regelmäßige Beschickung des* ***Versöhnungsfriedhofs*** *zu sichern* […] (GGU, S. 139) Das Lexem *Versöhnungsfriedhofs* stellt eine semantische Themaentwicklung dar.

[…] *das* ***Sekretariat*** *so früh einzurichten* […] (GGU, S. 139) Bei dem *Sekretariat* wird aufgrund des Kontextes ein Bezug auf die FG als semantische Themaentwicklung hergestellt.

Die Zahl der ***Beerdigungswilligen*** *wächst*. (GGU, S. 139) Die *Beerdigungswilligen* sind durch den Kontext als Kunden der FG zu erkennen, so kommt eine semantische Themaentwicklung zustande.

[…] *neben vielen* ***Kleinspenden****, beachtliche* ***Beträge*** […] *auf dem seit Anfang Juni eröffneten* ***Spendenkonto*** […] (GGU, S. 139) Die Lexeme *Kleinspenden*, *Beträge* und *Spendenkonto* beziehen sich inhaltlich auf den Referenzbereich der FG, denn sie gehören zu dem finanziellen Bereich der Gesellschaft. Auf diese Weise werden sie als Hinweise der semantischen Themaentwicklung bewertet.

Später kam es zu ***Überführungen*** *aus Übersee*. (GGU, S. 141) Die FG vermittelt als angebotene Dienste auch die *Überführungen*, dieser inhaltliche Bezug wird im Text als semantischer Themahinweis realisiert.

102 Direkte Rede der Romanfigur Brakup.

*Doch einen Großteil der **Korrespondenz** konnte er seiner **Sekretärin** überlassen* […] (GGU, S. 141) Bei dem Beleg *Korrespondenz* wird durch den Kontext angezeigt, dass es sich um (Geschäfts-)Korrespondenz in Bezug auf die FG handelt. So bildet *Korrespondenz* zusammen mit *Sekretärin* (als Personal der FG) Hinweise auf die semantische Themaentwicklung.

[...] *Kindheitserinnerungen« seiner **Sekretärin**, die sich allerdings nicht in die **Kartei** der im Computer versammelten **Beerdigungswilligen** einreihen wollte*. (GGU, S. 141) Die Lexeme *Sekretärin* (Referenz auf das Personal), *Kartei* (Referenz auf eine Datenbank der Kontakte an die Kunden der FG) und *Beerdigungswilligen* (Referenz auf Kunden der FG) stellen semantische Themaentwicklungshinweise dar.

[…] *überläßt ihr das **Sekretariat*** […] (GGU, S. 142) Das *Sekretariat*, im Sinne einer Abteilung der FG, wird als semantische Themaentwicklung bezeichnet.

*Und dennoch wird die **Friedhofsgesellschaft** unter der veränderten Marktsituation kaum zu leiden haben.* (GGU, S. 142) Hier liegt eine Themabeibehaltung durch Rekurrenz von *Friedhofsgesellschaft* vor.

Während eines Picknicks sagt Alexandra, dass sie aufhören sollen, solange alles noch gut läuft und es noch schön ist: *»Bloß weil wir haben **Idee** gehabt?«* (GGU, S. 144) Das Lexem *Idee* referiert auf die FG als Substitution im Rahmen der Themaentwicklung.

Der Betrieb der Gesellschaft wird zwar nicht unterbrochen, aber nach und nach entfernt sich die Friedhofsgesellschaft von der ursprünglichen Idee.

*Drei **Begräbnisse** standen auf dem Programm*, […] (GGU, S. 145) Die *Begräbnisse* gehören zu den angebotenen Dienstleistungen der FG und werden als semantische Themaentwicklung interpretiert.

[…] *was der* ***Gesellschaftsvertrag*** *zuließ.* (GGU, S. 145) Das Weltwissen stellt den Zusammenhang zwischen der FG und dem *Gesellschaftsvertrag* her, und so entsteht ein semantischer Themaentwicklungshinweis.

Mit der ***Beerdigung*** *nach katholischem Ritual* […] (GGU, S. 145) Die *Beerdigung* (u. U. auch das ‚katholische Ritual') wird von der FG organisiert. Dieser inhaltliche, durch den Kontext bestätigte Zusammenhang begründet eine semantische Themaentwicklung.

Den Fotos nach […] *hätten die polnischen* ***Sargträger*** *auch deutsche sein können. Mittlerweile waren, außer den* ***Totengräbern****, zwei* ***Friedhofsgärtner*** *fest angestellt. Zur Großen Allee hin saß im Backsteinhaus eine* ***Aufsichtsperson****,* [...] (GGU, S. 145) Die Lexeme *Sargträger*, *Totengräbern*, *Friedhofsgärtner* und *Aufsichtsperson* bezeichnen das bei der FG angestelltes Personal, der Bezug wird durch den Kontext bestätigt; diese Belege werden als semantische Themaentwicklungshinweise bezeichnet.

Gerne hätte die Brakup diesen vorerst noch ruhigen ***Posten*** *bezogen. Mit Mühe hatte ihr Jerzy Wróbel die Unvereinbarkeit eines zum* ***Friedhof*** *gehörenden* ***Arbeitsplatzes*** *mit ihrem* ***Sitz in der Aufsicht*** *erklärt, doch erst als er der Brakup den hohen Grad ihrer Verantwortung als* ***Sprecherin*** *einer Minderheit deutlich gemacht hatte* […] (GGU, S. 146) Die hier vertretenen Belege referieren auf die FG als semantische Themaentwicklung: die Umschreibung der Arbeitsmöglichkeiten bei der FG mit *Posten*, *Arbeitsplatzes* und das Erwähnen der Position *Sprecherin*, sowie der von der FG betriebene *Friedhof* und *Aufsicht* als ein Organ der FG und schließlich der *Sitz* als Position in der Aufsicht. Dieselbe Klassifikation erfolgt auch bei den folgenden Belegen: […] ***Aufsicht*** *sain auffem* ***Friedhof*** ...[103] (GGU, S. 146)

Im weiteren Textverlauf wird der Betrieb beschrieben:

Im übrigen ist der ***Friedhof*** *Aufgabe genug. Unsere* ***Idee*** *weist nach vorne,* […] (GGU, S. 148) Der Beleg *Friedhof* stellt den von der FG betriebenen Friedhof dar, so wird er als semantischer Themaentwicklungshinweis interpretiert. Das Lexem *Idee* wird

103 Direkte Rede der Romanfigur Brakup.

durch den Kontext als ‚die Idee der Gründung der Friedhofgesellschaft', identifiziert, die durch Substitution als Themaentwicklung auf die FG referiert.

[…] *konnten nur die bisher üblichen* ***Erdbestattungen*** *gemeint sein; schon Ende Juli wurde es notwendig, auf dem* ***Versöhnungsfriedhof*** *ein Urnenfeld anzulegen, weil eine wachsende Zahl »****Beerdigungswilliger****« auf* ***Feuerbestattung*** *Wert legte.* (GGU, S. 148) Die untersuchten Lexeme *Erdbestattungen* und *Feuerbestattung* beziehen sich auf die FG als angebotene Dienstleistungen; *Versöhnungsfriedhof* ist der von der FG betriebene Friedhof, und *Beerdigungswilliger* referieret auf die FG als ihre Kunden. Alle diese Belege werden als Hinweise auf die semantische Themaentwicklung bewertet.

[…] *verlangten* ***Einäscherung*** *und verzichteten auf christliche* ***Begräbnisrituale****. Ohne sich Atheisten zu nennen, wollten die* ***Antragsteller*** […] *»nur eine schlichte* ***Beisetzung*** *ohne Pfarrer und Reden am Grab«. Für diese Wünsche werden außerdem die niedrigen Kosten, insbesondere bei* ***Urnenüberführungen****, gesprochen haben,* […] (GGU, S. 148) Die *Einäscherung*, *Beisetzung* und *Urnenüberführungen* stellen Dienstleistungen der FG dar, u. U. kann angenommen werden, dass die FG auch die christlichen *Begräbnisrituale* vermittelt, so wird der inhaltliche Zusammenhang zur FG hergestellt. Auch *Antragsteller* haben als Kunden einen Bezug auf die FG. Die Belege dieses Textabschnitts werden als Hinweise auf die semantische Themaentwicklung zusammengefasst.

[…] *im westlichen Bereich des* ***Versöhnungsfriedhofs*** *angelegt*, […] (GGU, S. 148) Der *Versöhnungsfriedhof* ist der von der FG betriebene Friedhof, der inhaltliche Zusammenhang bestätigt die semantische Themaentwicklung.

Bei den ***Urnenbeisetzungen*** *war der Kreis der Leidtragenden kleiner,* [...] (GGU, S. 149) Bei *Urnenbeisetzungen* findet die semantische Themaentwicklung durch die angebotenen Dienste der FG statt.

Solche Rückgriffe sollten schon bald den ***Aufsichtsrat*** *der* ***Deutsch-Polnischen Friedhofsgesellschaft*** *in Schwierigkeiten bringen;* […] (GGU, S. 149) Der Beleg *Aufsichtsrat* referiert als ein Organ der FG auf die Friedhofsgesellschaft und wird als

semantische Themaentwicklung interpretiert. Das Lexem *Friedhofsgesellschaft* stellt eine Themabeibehaltung durch Rekurrenz dar.

[…] *mit Hinweisen auf die* ***Friedhofsordnung*** […] (GGU, S. 149) Durch das Weltwissen wird begründet, dass eine Gesellschaft auch eine Ordnung hat, so referiert *Friedhofsordnung* als eine eingeführte Regulierung oder Vorschrift auf die FG und bildet einen semantischen Entwicklungshinweis.

Nach Zahlung eines zinslosen Kredits, den der ***Aufsichtsrat*** […] *genehmigte, ließ sich auf dem Gelände des ehemaligen Krematoriums ein Steinmetzbetrieb nieder, der bald für den* ***Versöhnungsfriedhof*** *auf Vorrat zu arbeiten begann*. (GGU, S. 149) Die Lexeme *Aufsichtsrat* (als ein Organ der Gesellschaft) und *Versöhnungsfriedhof* (als der von FG betriebene Friedhof) werden als semantische Themaentwicklungshinweise interpretiert.

Reschke achtete streng darauf, daß kein Stein den Vorschriften der ***Friedhofsordnung*** *querstand.* (GGU, S. 150) Die *Friedhofsordnung* im Sinne einer Regulierung oder Vorschrift referiert auf die FG als semantischer Thema-entwicklungshinweis.

Der ***Aufsichtsrat*** *mußte bemüht werden,* […] (GGU, S. 151) Das Lexem *Aufsichtsrat* wird im Sinne eines Organs der FG als semantische Entwicklung interpretiert.

Auf dem Friedhof kommt es zu Diebstählen von Kränzen und Blumen, und der Aufsichtsrat muss diese Situation bewältigen:

[…] ***der Deutsch-Polnischen Friedhofsgesellschaft*** *stand Streit bevor. Die Frage, ob das* ***Versöhnungsfriedhof*** *heißende* ***Parkgelände*** *[...] eingezäunt werden sollte, wurde früh, gleich nach Gründung der* ***Friedhofsgesellschaft*** *gestellt, doch als nicht dringlich vertagt.* (GGU, S. 152)

Beide Belege von *Friedhofsgesellschaft* (auch innerhalb der Wortgruppe *Deutsch-Polnischen Friedhofsgesellschaft*) stellen Hinweise auf Themabeibehaltung durch Rekurrenz dar. Der *Versöhnungsfriedhof* bezeichnet den von der FG betriebenen Friedhof, und das *Parkgelände* bezeichnet das Grundstück, das durch den Kontext als

die Lage des Friedhofs der FG verstanden wird – diese beiden Belege werden als Hinweis auf die semantische Themaentwicklung interpretiert.

[…] *faßte der gerade noch beschlußfähige* ***Aufsichtsrat*** […] *nach zu kurzer Debatte* […] *den Beschluß, das* ***Friedhofsgelände*** *durch einen Zaun zu schützen und obendrein* ***Nachtwächter*** *einzustellen, selbstverständlich auf* ***Kosten*** *der* ***Friedhofs-gesellschaft***. (GGU, S. 152)

Der *Aufsichtsrat* tritt als ein Organ der FG auf. Das *Friedhofsgelände* bezieht sich auf den von der FG betriebenen Friedhof. Der *Nachtwächter* wird von der FG eingestellt und gehört zu dem Personal. Der Kontext bestätigt, dass es sich bei Kosten um die zu bezahlende Kosten der FG handelt. Diese vier Belege referieren auf die FG und werden als semantische Themaentwicklungshinweise interpretiert. Der Ausdruck *Friedhofsgesellschaft* stellt einen Themabeibehaltungshinweis durch Rekurrenz dar.

Jetzt rächte sich, daß die Piątkowska und Reschke als geschäftsführende ***Gesellschafter*** *keine Stimme im* ***Aufsichtsrat*** *hatten.* […] *Sogar Wróbel stimmte für den Zaun*. (GGU, S. 153)

Das Weltwissen zeigt, dass die *Gesellschafter* als Teilhaber an dem Unternehmen[104] auf die FG referieren. Der *Aufsichtsrat* wird als ein Organ der FG interpretiert. Beide Lexeme werden als Hinweise auf die semantische Thema-entwicklung bewertet.

Der Aufsichtsrat beschließt, das Problem mit den Diebstählen durch das Einzäunen des Friedhofs zu lösen, aber der Zaun ruft negative Reaktionen hervor:

[…] *Bauelemente der jüngst hinfällig gewordenen Berliner Mauer zollfrei nach Polen einführen, um sie zum Schutz des deutschen* ***Friedhofs*** *abermals aufzustellen*. (GGU, S. 153) Der Ausdruck *Friedhofs* wird innerhalb seines Kontextes als der von der FG betriebene Friedhof identifiziert und als semantische Entwicklung interpretiert.

104 Die Definition von ‚Gesellschafter' vgl. Duden – Deutsches Universalwörterbuch, 6. Aufl. Mannheim 2006 [CD-ROM].

Denn täglich konnten sechs bis zehn ***Erdbestattungen*** *gezählt werden.* (GGU, S. 153) Das Lexem *Erdbestattungen* gehört zu den von der FG angebotenen Diensten und wird als semantische Themaentwicklung bezeichnet.

Der ***Friedhof*** *florierte* […] (GGU, S. 154) Der *Friedhof* im Sinne eines von der FG betriebenen Friedhofs wird als semantische Entwicklung interpretiert.

Als schließlich Protestveranstaltungen […] *die Friedhofsruhe störten, zog der* ***Aufsichtsrat*** *nach fernmündlicher Beratung seinen Beschluß zurück:* […] (GGU, S. 154) Bei dem Lexem *Aufsichtsrat* liegt semantische Themaentwicklung vor.

Der ***Aufsichtsrat*** *erklärte öffentlich sein Bedauern, und die Journalisten verloren den Spaß an der Sache. Wenngleich die Blumen- und Kranzdiebstähle, trotz* ***Nachtwächterdienst****, nie ganz aufhörten,* […] (GGU, S. 154)

Der *Aufsichtsrat* stellt ein Organ der FG dar. Der *Nachtwächterdienst* ist eine von der FG angebotene Dienstleistung, durch die das Friedhofsgelände in Sicherheit vor Diebstählen zu halten ist. Die beiden Belege in diesem Textabschnitt werden als semantische Themaentwicklung klassifiziert.

Die Situation mit den Diebstählen auf dem Friedhof hat sich beruhigt und der Betrieb geht ungestört weiter:

Auf ***Beerdigungen*** *sind sie ab Ende August nur selten gegangen. Man überließ das Wróbel und Erna Brakup, die keine* ***Beisetzung*** *und im Hevelius keinen Leichenschmaus verpaßten*. (GGU, S. 156)

Der Kontext erklärt die *Beerdigungen* und *Beisetzung* als die von der FG angebotenen Dienste. Beide diese Ausdrücke werden als Hinweise auf die semantische Entwicklung interpretiert.

Der ***Versöhnungsfriedhof*** *war jetzt ungestört in Betrieb*. […] *Belieferung des* ***Versöhnungsfriedhofes***. (GGU, S. 157) Das Lexem *Versöhnungsfriedhof* wird in beiden Fällen durch den Kontext als der von der FG betriebene Friedhof identifiziert und als semantische Entwicklung bezeichnet.

*Der **Kontostand** der **Friedhofsgesellschaft** bewies jedoch, daß die gespannte Weltlage dem **deutsch-polnischen Versöhnungswerk** nicht abträglich war.* (GGU, S. 157) Der Kontext zeigt den Bezug von *Kontostand* auf die FG, es handelt sich um ein Konto der FG, und auf diese Weise wird ein semantischer Themaentwicklungshinweis zum Ausdruck gebracht. Der Beleg *Friedhofsgesellschaft* (auch innerhalb der Wortgruppe *der Kontostand der Friedhofsgesellschaft*) zeigt die Themabeibehaltung durch Rekurrenz. Das Lexem *Versöhnungswerk* wird als Referenz auf die FG betrachtet und als Themaentwicklung durch Substitution interpretiert.

*Trotz erheblicher **Neben- und Unterhaltskosten** schnellten die **Konten** in unerwartete Höhe*. […] *die anfallenden **Überführungs- und Nutzungskosten*** […] *nur unerheblich das **Kapital** minderten*. (GGU, S. 157)

Der Kontext zeigt, dass die *Neben- und Unterhaltskosten* und *Überführungs- und Nutzungskosten* die von der FG zu bezahlenden Kosten darstellen, die *Konten* zum Besitz der FG gehören und das *Kapital* für die finanziellen Mittel der FG steht. Diese Ausdrücke referieren auf die FG als semantische Entwicklung.

*Nicht zu reden vom **Spendenkonto**, mit dessen Hilfe die **Begräbniskosten** der Bedürftigen beglichen wurden* […] (GGU, S. 157)

Durch den Kontext wird erschlossen, dass das *Spendenkonto* von der FG besessen wird. Die *Begräbniskosten* sind solche Kosten, die der FG zu bezahlen sind. Durch den Kontext werden diese Belege als semantische Entwicklung bezeichnet.

[…] *ein letzter Platz auf dem **Versöhnungsfriedhof** sicher sein, die **DPFG** wirkte gemeinnützig*. (GGU, S. 157)

Der Beleg *Versöhnungsfriedhof* stellt den von der FG betriebenen Friedhof dar und realisiert so die semantische Themaentwicklung. Die Abkürzung *DPFG* enthält in ihrer voller Form den Hinweis auf die Themabeibehaltung durch Rekurrenz des Ausdrucks ‚Friedhofsgesellschaft'.

*Trotz der günstigen **Kontostände** muß gesagt werden, daß ein Drittel des **Kapitals** samt Zinsen der immer noch ruhenden litauischen **Vertragskomponente** vorbehalten war;* […] (GGU, S. 158)

Der Kontext vermittelt die Information, dass *Kontostände* als die finanziellen Mittel auf dem Konto der FG einen Zusammenhang mit der FG haben. Ebenso bezeichnet das *Kapital* die Finanzen der FG. Die erwähnte *Vertragskomponente* weist auf den geplanten Friedhof in Wilno hin, der von der FG betrieben werden soll. Diese auf dem Kontext beruhenden Referenzen auf die FG werden als semantische Themaentwicklungshinweise zusammengefasst.

Die Komponente in Wilno: *Dem* ***Extrafriedhof*** *für tote Polen,* […] (GGU, S. 158) Der *Extrafriedhof* ist einer der von der FG betriebenen Friedhöfe, der kontextuelle Bezug hilft den Beleg als semantische Themaentwicklung erkennen.

Reschke schließt mit Chatterjee ein Abkommen über Kapitalbeteiligung im Rikschabetrieb:

[...] *aber die dreißigprozentige* ***Kapitalbeteiligung der Deutsch-Polnischen Friedhofsgesellschaft*** *an S. Ch. Chatterjees Fahrradrikschaproduktion ist erst anläßlich einer viel später einberufenen Sitzung der* ***Aufsicht*** *ans Licht gekommen.* […] *Der* ***Kontostand*** *der* ***Friedhofsgesellschaft*** *erlaubte die Transaktion.* (GGU, S. 163)

Bei dem Beleg *Kapitalbeteiligung* wird die Referenz auf die FG durch den Kontext als eine von der FG durchgeführte Investition sichtbar und als semantische Themaentwicklung interpretiert. Das Lexem *Friedhofsgesellschaft* (bzw. beide Belege in diesem Textabschnitt) ist eine Themabeibehaltung durch Rekurrenz. Bei *Aufsicht* kommt semantische Themaentwicklung zustande. Der Kontext des Lexems *Kontostand* gibt direkt an, dass es sich um ein Konto der FG handelt und realisiert auf diese Weise einen semantischen Themaentwicklungshinweis.

[…] *wie sie tagtäglich auf unserem* ***Versöhnungsfriedhof*** *stattfindet,* […] (GGU, S. 164) Der *Versöhnungsfriedhof* wird im Sinne des von der FG betriebenen Friedhofs als Hinweis auf semantische Themaentwicklung interpretiert.

Das vierte Kapitel wird mit einem Gespräch mit Chatterjee abgeschlossen. Nachdem die Hinweise zu Einführung, Beibehaltung und Entwicklung des Teilthemas beobachtet wurden, wird die Aufmerksamkeit nun auf den Abschluss des Teilthemas gerichtet. Dem Textinhalt wurde entnommen, dass das Teilthema ‚Friedhofsgesellschaft'

im ganzen Romantext vertreten ist, und so wird nach dem Abschluss dieses Teilthemas im letzten, im siebten Kapitel gesucht.

Um in den Themaabschluss und den dazu gehörenden Kontext einzuführen, soll noch kurz auf die Handlung im siebten Kapitel eingegangen werden: Reschke hat Piątkowska einen Heiratsantrag gemacht. Das Paar tritt endgültig aus der Gesellschaft aus. Die FG entwickelt sich weiter, es wird ein neuer Sitzungsraum im Altstädtischen Rathaus genutzt, die Projekte der neuen Aufsichtsratmitglieder werden realisiert. Reschke muss sein Finanzgebaren vor dem Aufsichtsrat erklären. Später wird die Hochzeit von Reschke und Piątkowska gefeiert. Nach der Hochzeit kommt die Idee auf, eine Chronik der FG zu schreiben.

Im letzten Abschnitt des Romans tritt das Paar eine Hochzeitsreise nach Italien an und nach dem Besuch einiger italienischer Städte steht Neapel auf dem Plan.

*Das **Ende**, falls es ein **Ende** gibt, steht fest*. (GGU, S. 298) Das Ende der Geschichte wird metakommunikativ, aus der Perspektive des Erzählers, angedeutet.

Während eines Autounfalls sterben Reschke und Piątkowska und werden auf einem Dorffriedhof in Italien begraben, denn: *Sie waren gegen **Umbettung***. (GGU, S. 299) Das Substantiv *Umbettung*, als ein Projekt der FG, referiert auf die Friedhofgesellschaft und wird als ein Hinweis auf die semantische Themaentwicklung interpretiert.

Im Text kommen keine weiteren Hinweise auf die Friedhofsgesellschaft vor. In den letzten Sätzen wird nur noch pronominal auf das Paar hingewiesen: ***Sie** liegen gut da. Laßt **sie** da liegen*. (GGU, S. 299) mit dem Tod des Paares endet also nicht nur die Liebesgeschichte, sondern auch die Handlung über die FG wird abgeschlossen.

Der Themaabschluss wird hier durch außersprachliche Mittel realisiert, weil sich die Themaabschlusshinweise auf die Friedhofsgesellschaft mit dem Ende der textuellen Einheit decken, und das Ende des Romans setzt die Grenze für alle weiteren Themahinweise.

Da nicht der komplette Text untersucht wurde, sondern nur ausgewählte Kapitel, wird abschließend keine zusammenfassende statistische Übersicht der Belege (wie bei dem Referenzbereich ‚Engel‘) angeführt.

5.5 Zusammenfassung

In diesem Kapitel wurde versucht, die Analyse der Teilthemaentwicklung im Gesamttext (dies ist hier der Roman ‚Unkenrufe' von Günter Grass in deutscher Fassung) durchzuführen.

Das erste Unterkapitel stellt die einzelnen Themahinweise vor. Das Thema des Textes drückt aus, wovon der Text handelt. Außer dem Gesamttextthema werden auch Neben- oder Teilthemen unterschieden, die sich auf bestimmte Textabschnitte beziehen. Die Frage ist, wie auf das Teilthema im Text hingewiesen wird und wie sich diese Teilthemen noch weiter teilen lassen. Bei einem Teilthema kann seine Einführung, Beibehaltung, Entwicklung und sein Abschluss im Text beobachtet werden.

Das zweite Unterkapitel führt eine Analyse an, die den Vorgang für die Untersuchung im Gesamttext einleiten soll, und soll dadurch helfen, die relevanten Schritte der Analyse im Gesamttext zu formulieren Satz genauer formulieren. Als Ausgangspunkt wurde stellvertretend ein Teilthema ausgewählt, das die Grenze eines Kapitels nicht überschreitet. Es wurde mit der Suche nach den rekurrenten Belegen angefangen, denn diese können einfach mit der gewöhnlichen Suchfunktion, die ein Computer anbietet, gefunden werden. Die folgende Suche nach den Belegen hängt in einem nicht grammatisch annotierten Korpustext von dem Benutzer ab. Weitere Belege können mit Anwendung der Regeln der behandelten Wortschatzgruppierung (vgl. Unterkapitel „4.1.2 Wortschatzgruppierung") oder durch die Mittel der Kohäsion und Kohärenz und Wortbildung (vgl. Unterkapitel „4.2 Textkonstitution") identifiziert und gesammelt werden. Im Allgemeinen beginnt die Suche nach den zusammenhängenden Belegen bei der Ausdrucksseite der Lexeme und wird mit der Berücksichtigung der Inhaltsseite ergänzt. Die gesammelten Belege wurden im Sinne der Themahinweise untersucht, um die einzelnen Phasen der Themaentwicklung im Text beschreiben zu können.

Das abschließende Unterkapitel befasst sich mit der Entwicklung des Teilthemas im Gesamttext. Das Teilthema, als ein untergeordnetes Element des Textthemas, wurde unter Verwendung der Wortformenliste identifiziert. Bei der Bestimmung der relevanten Bereiche wird die Vorkommensfrequenz der Einträge im Gesamttext beurteilt. Es hat sich gezeigt, dass anhand der Angaben in der Frequenz-Wortformenliste der für die weitere Textuntersuchung bedeutende Referenzbereich die ‚Friedhofsgesellschaft' ist.

Die Hinweise auf das Teilthema werden anhand der Beziehungen zwischen den Belegen im Text untersucht. Diese Beziehungen werden nicht nur durch die Zugehörigkeit der Lexeme zueinander innerhalb der behandelten Wortschatzgruppierungsarten festgelegt. Eine bedeutende Rolle spielt auch der Kontext, und so muss auf die Regeln der Kohäsions- und Kohärenzbeziehungen geachtet werden. Die Untersuchung ging größtenteils von Substantiven aus, ein Untersuchungsvorschlag wäre, auch andere Wortarten bzw. auch Raum- oder Zeitreferenzen bei der Analyse zu berücksichtigen.

Schlussfolgerungen

Die vorliegende Arbeit ist in zwei elementare Teile, den technischen und den analytischen, gegliedert.

Im technischen Teil wird die Aufbereitung der Korpustexte behandelt. Das erste Kapitel hat das Ziel, die Korpuslinguistik vorzustellen. Die Korpuslinguistik stellt kein eigenes sprachwissenschaftliches Gebiet dar, sondern wird als eine Methodologie verstanden, welche die Untersuchung der sprachlichen Mittel durch ein Sprachkorpus als Quelle authentischer Sprachdaten unterstützt. Um eine beliebige Textsammlung von einem Sprachkorpus zu unterscheiden, müssen Kriterien zur Definition eines Korpus erstellt werden. Das Ziel dieser Arbeit ist es, ein bilinguales Korpus zu erstellen, das als eine Sammlung schriftlicher sprachlicher Äußerungen, die im Computer gespeichert und maschinenlesbar sind, verstanden wird. Die Quelldaten sind mit Metadaten versehen, die die bibliografischen Angaben, Struktur und Segmentierung des Ausgangstextes kodieren. Die Beschreibung in dem entsprechenden Kapitel listet die Kriterien der heute verwendeten Sprachkorpora auf und stellt ausgewählte Korpusarten vor. Es gibt nicht nur maschinenlesbare Korpora der geschriebenen Sprache, es können auch kleinere Sprachanalysen an gedruckten oder handgeschriebenen Texten durchgeführt oder Ton- oder Videoaufnahmen untersucht werden. Diese und weitere ähnliche Korpora werden in dieser Arbeit nicht behandelt. Bei den Korpora können weitere Eigenschaften und Kriterien der Aufteilung der Korpusarten auftreten, wie: Ein- oder Mehrsprachigkeit, Größe, Annotation oder geschichtlicher Zeitraum. Das hier bearbeitete DeuCze-Korpus enthält Texte in zwei Sprachen (Deutsch und Tschechisch); die Daten werden mit Metadaten zur Strukturbeschreibung versehen, enthalten jedoch (noch) keine grammatische Annotation.

Das DeuCze-Korpus ist dank der Zusammenarbeit der germanistischen Abteilung der Universität Opava und des Lehrstuhls für deutsche Sprachwissenschaft der Universität Würzburg entstanden. Dieses Korpus besteht aus belletristischen Texten, die in zwei unterschiedlichen Sprachen verfasst worden sind. Es sind jeweils zwei deutsche Originaltexte mit tschechischen Übersetzungen und zwei tschechische Originaltexte mit deutschen Übersetzungen vertreten, das DeuCze-Korpus ist also ein Parallelkorpus. Im Rahmen dieser Arbeit wurden zwei Romane jeweils mit ihren Über-

setzungen digitalisiert:

- Jiří Kratochvil: *Nesmrtelný příběh aneb Život Soni Trocké-Sammlerové čili Román karneval.* Verlag Petrov, Brno 2005 und die Übersetzung ins Deutsche von Liedtke, Kathrin und Vagadayová, Milka: *Unsterbliche Geschichte oder Das Leben von Sonja Trotzkij-Sammler oder Karneval.* Verlag Amman, Zürich 2000.
- Thomas Brussig: *Am kürzeren Ende der Sonnenallee.* Fischer Taschenbuch Verlag, Frankfurt am Main 2001 und die Übersetzung ins Tschechische von Zoubková, Jana: *Na kratším konci ulice.* Verlag Odeon, Praha 2001.

Das Korpus gehört zu den kleinen Korpora, zurzeit enthält es knapp 500 000 Wortformen[105].

Das zweite Kapitel enthält eine Beschreibung der Vorbereitungsschritte für die Digitalisierung der Texte für das Korpus. Die Texte werden typografisch analysiert, um ihre Struktur zu erschließen, d. h., aus den Primärquellen wurden Informationen zur Textgestaltung gewonnen, und dadurch konnte das System der Tags entworfen werden. Die Tags sind Marken einer Auszeichnungssprache, hier XML, mit deren Hilfe dem Text unterschiedliche Metainformationen (wie die Markierung der Kapitel, Absätze, Zeilenumbrüche usw.) hinzugefügt werden. Die Ziele der Digitalisierung sollten vor dem eigentlichen Scannen möglichst genau bestimmt werden, denn die Qualität der gescannten Bilder ist für die künftige Verwendung wichtig. Es muss also unterschieden werden, ob die Bilder am Bildschirm angezeigt werden oder zu weiterer Bearbeitung dienen sollen. Die Bücher wurden seitenweise gescannt, und jede Seite wird als eine selbstständige Bilddatei gespeichert. In diesem Projekt wird mit zwei Typen von Bildern gearbeitet; einerseits sind es die sog. Masterdateien, die eine hohe Qualität haben und für die Archivierung vorgesehen sind; andererseits werden von den Bildern kleinere Kopien abgeleitet, sog. Derivative, die bei einer noch ausreichenden Qualität weniger Speicherplatz benötigen und eine kürzere Übertragungszeit über das Internet in Anspruch nehmen. Konkrete Scaneinstellungen und Speicherformate wurden in einem eigenen Schritt mit Probescans getestet und optimale Werte für das Scannen und Speichern der Bilder festgelegt.

105 Diese Angabe stammt aus der DeuCze-Website ‚deucze.org' bzw. http://www.deucze.germanistik.uni-wuerzburg.de, zit.: 23. 6. 2010.

Die Bilddigitalisierung wird in einem eigenen Unterkapitel thematisiert. Diese Beschreibung geht größtenteils von allgemeinen theoretischen Erfahrungen aus, die dann durch die Probescans getestet und konkret angewendet werden. Für die weitere Bearbeitung ist die Anzahl der enthaltenen Farben der gescannten Bilder entscheidend. Bei bitonalen (schwarz-weißen) Grafiken können manche Informationen verloren gehen, z. B. können diakritische Zeichen nicht ausreichend abgebildet sein. Durch die Verwendung von Graustufen können detailliertere Abstufungen abgebildet werden, so dass der dargestellte Text auf dem gescannten Bild besser lesbar ist. Das Scannen in Farben kommt bei den hier bearbeiteten Textseiten nicht zum Einsatz, obwohl alle drei erwähnten Scanmodi durch Probescans getestet wurden. Ein weiteres Kriterium für das Bewerten der Bildqualität ist die Auflösung, die die Anzahl der auf die Länge von einem Zoll enthaltenen Bildelemente (Pixel) angibt. Durch höhere Auflösung wird bessere Detaildarstellung erreicht, und die Ergebnisse der optischen Zeichenerkennung werden positiv beeinflusst. Unterschiedliche Werte der Auflösung wurden mit Probescans getestet. Beim Scannen muss darauf geachtet werden, dass die Buchseiten gerade auf das Scannerglas gelegt, in der richtigen Reihenfolge und vollständig gescannt werden. Mangelhaft gescannte Seiten müssen neu gescannt werden.

Im Unterkapitel „2.4 Textscannen und OCR-Textbearbeitung“ werden zunächst die Scanner-Einstellungen unter Berücksichtigung der hier bearbeiteten Texte behandelt. In diesem Abschnitt handelt es sich um Vorbereitungen, die eigentliche Bearbeitung der Bücher wurde später durchgeführt – nachdem die Ergebnisse der Probescans gewonnen wurden. Zu den erwähnten Kriterien gehört auch der Kontrast, denn durch entsprechende Werte können bessere OCR-Ergebnisse erreicht werden. Die nächste Verarbeitungsphase besteht in der optischen Zeichenerkennung, die das Programm ABBY FineReader durchführt. Die Pixel-Grafiken werden in editierbare Textdateien umgewandelt, dabei wird ein gutes Resultat von mehreren Maßnahmen unterstützt. Neben den Scaner-Einstellungen können im OCR-Programm auch andere Einstellungen für das jeweilige Projekt angepasst werden, wie die Dokumentsprache oder die Option, die es ermöglicht, dass das Programm auf neue Muster trainiert wird. Der entstandene Text wird im gewählten Format gespeichert, hier wird das MS Word-Format (Dateierweiterung: .doc) gewählt, weil auf diese Weise auch die Formatierung (z. B. Kursiv- oder Kapitälchenschrift) erhalten bleibt. Anschließend werden die gewonnenen Texte in

mehreren Schritten daraufhin kontrolliert, ob Layout und Textinhalt der Vorlage entsprechen.

Im abschließenden Teil dieses Kapitels wird die Durchführung der Probescans thematisiert, um zu testen, wie sich die Farb- und Auflösungswerte zusammen mit den zwei ausgewählten Speicherformaten auf das OCR-Ergebnis auswirken. Die Ergebnisse der Proben geben die optimalen Scaneinstellungen an. Für das Scannen der Archivdateien, der Masters, werden die Einstellungen für Farbtiefe: grau, für die Auflösung: 600 dpi und das Speicherformat: TIFF verwendet. Das TIFF-Speicherformat ermöglicht es, die Bilder unkomprimiert zu speichern, und es können auch Informationen über die jeweilige Grafik in ihrem Header gespeichert werden. Für die Benutzeroberfläche des Korpus werden kleinere Kopien, sog. Derivative, gebildet – sie werden auch in Graustufen angefertigt, doch die Auflösung wird auf 150 dpi gesenkt und das Speicherformat ist JPG. Die in diesem Speicherformat komprimierten Bilder benötigen weniger Speicherplatz und kürzere Übertragungszeiten im Internet.

Die Aufbereitung der Texte für das Korpus wird im dritten Kapitel abgeschlossen, indem die Texte mit der XML-Markierung ergänzt wurden. Die zusätzlichen Informationen über die typografische Gestaltung und über die bibliografischen Angaben werden im Text mit den Mitteln der Auszeichnungssprache XML (‚eXtensible Markup Language‘) realisiert. Das gewählte Format XML und seine Eigenschaften (wie die Kodierung Unicode UTF-8) ermöglichen es, dass die Dateien nicht an eine bestimmte Plattform (getestet wurde MS Windows und Ubuntu Linux) oder proprietäre Software gebunden sind. Eine XML-Datei muss dabei zwei weitere Kriterien erfüllen, dies sind die Wohlgeformtheit und die Validität. Die Datei ist dann wohlgeformt, wenn die hierarchische XML-Struktur, die sich in der Verschachtelung der Elemente widerspiegelt, eingehalten wird. Die Validität bedeutet, dass die Elemente und ihre Attribute nach bestimmten Regeln kombiniert werden. Diese Verwendungsregeln werden in einem Schema für das jeweilige Dokument definiert. Bei der Verwendung der Elemente wurde von den Regeln der Text Encoding Initiative (in der Version P5) ausgegangen. Bei der Aufbereitung der Texte wird ein großer Wert auf Standards im Bereich der elektronischen Datenverarbeitung gelegt, denn auf diese Weise können die erstellten Dokumente mit anderen Projektteilnehmern problemlos ausgetauscht werden,

und es wird auch dafür gesorgt, dass die Daten auch nach Langzeitspeicherung immer noch verwendet werden können.

Das System der hier verwendeten Tags wird im anschließenden Unterkapitel vorgestellt. Es besteht aus einer Aufzählung und Beschreibung der einzelnen Tags, wie sie im deutschen Kratochvil-Text vorkommen. Dieser Text wurde als Beispiel gewählt, weil er von den hier bearbeiteten Texten die meisten zu markierenden Phänomene enthält. Im Vergleich dazu enthält der Text von Brussig z. B. keinen Brief. Die Tagstruktur betrifft nur den Textteil des Romans, der Dokument-Header wird in der Beschreibung nicht näher bearbeitet. Die obligatorischen Teile werden in den Guidelines von TEI ausführlich behandelt.[106]

Die Auflistung der verwendeten Tags entspricht dem Layout des gedruckten Textes. Dies bedeutet, dass unterschiedliche Strukturen (wie Kapitel oder einzelne Sätze), die im Text vertreten sind, entsprechend markiert sind. Außerdem wird noch das `<segment>`-Element verwendet. Dieses Element enthält im Normalfall je einen Satz bzw. eine satzartige Texteinheit. Wird z. B. ein Satz als zwei Sätze übersetzt, dann stehen diese zwei Sätze der Zielsprache in einem Segment. Die gleiche Anzahl der Segmente im Ausgangstext und in seiner Übersetzung ermöglicht die synoptisierte Anzeige nach einzelnen Sätzen der beiden Texte auf der Korpus-Benutzeroberfläche.

Bei den die hierarchische Struktur bezeichnenden Elementen wird das Attribut `xml:id` eingefügt. Dieses Attribut stellt eine eindeutige Referenz im Gesamttext dar, indem es die konkrete Textstelle identifiziert (wie Kapitel, Absatz, Segment und Satz in einem Satz-Element und Textbezeichnung mit Abkürzung von Autorennamen, Titel, Erscheinungsjahr und Sprache).

Im Bereich der linearen Struktur werden die Seiten- und Zeilenumbrüche markiert und nummeriert und die formatierten Textpassagen mit entsprechenden Elementen markiert, die die Art der Formatierung (Kursive, Fettschrift oder Kapitälchen) angeben.

Das anschließende Unterkapitel beschreibt den konkreten Weg, wie die Romantexte in die XML-Datei transformiert werden. Da die Texte von des OCR-Programms als MS Word-Dateien gespeichert wurden, wurden die ersten Schritte (wie die

106 Vgl. The TEI Consortium. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Zugänglich unter WWW: http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf, zit.: 13. 7. 2009.

Markierung der formatierten Textabschnitte oder das Korrekturlesen) mit dieser Software durchgeführt. Abschließend wurde der Text als einfache Textdatei ohne Formatierung und in der Zeichenkodierung UTF-8 gespeichert.

Die XML-Markierung wird durch ein PHP-Skript durchgeführt. Das Ziel dabei war, bestimmte allgemeine Bearbeitungsschritte zu entwerfen, die ihre Anwendung auch für andere Korpustexte finden können. Diese Schritte sind zum Beispiel die Markierung der Absätze, die Identifizierung der Satzgrenze nach allgemeingültigen Kriterien oder automatisches Einfügen der Seiten- und Zeilenumbruchmarkierung. Für jede Textdatei gibt es jeweils ein selbstständiges Skript, weil konkrete Korrekturen durchgeführt werden müssen (wie z. B. Korrektur bei Abkürzungen, bei denen der Punkt keine Satzgrenze markiert). Nachdem alle Elemente eingefügt wurden, konnten die vorgesehenen Attribute mit ihren Werten ergänzt werden, wie die Nummerierung und Identifizierung mit `xml:id`-Element, mithilfe einer XSLT-Schablone. Der letzte Schritt bei der Erstellung der XML-Datei ist das Erstellen des Headers. Dieser Teil der Datei enthält neben den bibliografischen Angaben auch eine Übersicht über die verwendeten Elemente und die Beschreibung der Textgestaltung bei der Vorlage (z. B. Einzug der Absätze), oder es werden auch die Stellen im Text erwähnt, bei denen Satzfehler korrigiert wurden. Die Bearbeitung der XML-Dateien wurde mit der Validierung abgeschlossen. Die Datei wird gegen ein Schema validiert, um zu testen, ob die Elemente regelrecht verwendet (verschachtelt und kombiniert) wurden.

Die elektronische Edition der Korpustexte ermöglichte nicht nur, dass die gesuchten Belege schnell gefunden werden konnten, sondern die enthaltenen Tags unterstützten auch die durchgeführte Textuntersuchung. Die Segmentierung der Texte nach Kapiteln, Absätzen und Sätzen vermittelt Informationen über die bearbeitete Textstelle, deren relative Position im Gesamttext auf diese Weise schnell erfassbar wird. Die Attribute bei einzelnen Elementen vermitteln Referenzangaben, die beim Zitieren der Belege anzuführen sind und bei Bedarf können Attribute eingeschaltet werden, die die Anzahl der Sätze innerhalb der `<seg>`-Elemente angeben.

Das beschriebene Arbeitsverfahren bei der Digitalisierung und Bearbeitung der Texte kann auch bei Erweiterung des DeuCze-Korpus verwendet werden. Natürlich müssen immer die jeweiligen Texte mit ihren Abweichungen von diesem Schema betrachtet und entsprechende Anpassungen in den Skripts durchgeführt werden.

Die Quelldateien könnten auch mit der Wortartenmarkierung, also mit der grammatischen Annotation, angereichert werden. Dies ist aber nicht der Gegenstand dieser Arbeit und betrifft eher die zukünftigen Projektphasen. Im Bereich der Segmentierung der Texte wäre es als ein weiterer Aspekt sicherlich interessant zu untersuchen, ob es Regelmäßigkeiten und Gründe dafür gibt, ob und wie bestimmte Sätze bei ihrer Übersetzung geteilt oder verknüpft werden.

Mit dem Erstellen der Textdaten für das Korpus ist der technische Teil der Arbeit abgeschlossen, und im analytischen Teil wird auf die Textanalyse eingegangen. Das Ziel dieses Teils der Arbeit ist eine Untersuchung der Teilthemaentwicklung im Gesamttext.

Im vierten Kapitel werden die theoretischen Grundlagen und die Vorbereitungsschritte für die Textanalyse vorgestellt. Bei der Analyse wurde mit dem Korpustext *Unkenrufe* von Günter Grass gearbeitet. Die Untersuchung geht von der Annahme aus, dass das Gesamttextthema in weitere Neben- oder Teilthemen aufgeteilt werden kann. Der erste Schritt bei der Untersuchung war zu bestimmen, welche Referenzbereiche in diesem Text am bedeutendsten sind. Dies wurde mit einer Wortformenliste durchgeführt, in der die Einträge der Frequenz nach sortiert wurden. In dieser Wortformenliste wurden die Doppelformen in ihren Vollformen ausgeschrieben. Jede Zeile der Wortformenliste stellt jeweils einen Type dar. Die Einträge wurden so sortiert, dass bei der Beobachtung bestimmte allgemeine Kategorien ausgelassen wurden, wie: Funktionswörter (Artikel, Präpositionen, Konjunktionen, Hilfsverben…), Namen der Protagonisten (*Reschke*, *Alexandra*, *Brakup*, …) und andere nicht motivierte Wortformen bzw. auch Verwandtschaftsbeziehungen (*Paar*, *Ende*, *Witwe*, …). Diese Beobachtung hat gezeigt, dass die drei am häufigsten vertretenen Referenzbereiche *Friedhof*, *Gesellschaft* und *Idee* sind. Die Einträge in dieser sortierten Liste stellen die Isotopieknoten dar, die auf Subthemen im Text referieren.

Die Wortformenliste führt alle Belege ohne Kontext an, bei der vorgesehenen Analyse jedoch sind die Wortbeziehungen im Text grundlegend. Die Zusammengehörigkeit der Wörter spielt eine wesentliche Rolle auch in der Textgestaltung. Zuerst wurden die Wörter nach den Regeln der in früheren Phasen der Untersuchung behandelten Arten der Wortschatzgruppierung betrachtet. Diese Gruppierungen sind formorientiert (Wortfamilien – Zugehörigkeit mittels morphologischer Prinzipien), inhaltsorientiert (onomasiologisches Paradigma – inhaltliche Zugehörigkeit zu einem

Referenzbereich) oder eine Kombination aus Form- und Inhaltsorientiertheit (Wortfeld – sinnverwandte Wörter).

Die Beziehungen zwischen den Wörtern helfen nicht nur bei der Satzbildung, sondern sie greifen auch über die Satzgrenze hinaus und wirken bei Gestaltung des Textthemas. Für die weiteren Betrachtungen ist der Ausgangspunkt die text-konstituierende Funktion der zwischenwörtlichen Beziehungen. Auf diese Weise helfen sie in den Texten die Kohäsions- (Ausdrucksseite) und Kohärenzzusammenhänge (Inhaltsseite) herzustellen. Es handelt sich um die Wiederaufnahme eines Ausdrucks aus einem Vorgängersatz in einem (nicht immer unmittelbar folgenden) Nachfolgersatz. Die Mittel der Kohäsion, also die Topikrelationen, können bei der Teilthemauntersuchung angewendet werden, denn sie können sich mit Themahinweisen decken. Bei der Suche nach den Themahinweisen ist auch die semantische Progression der Kohärenz behilflich. Im Kontext spielen die Semrekurrenz (die textuelle Aktivierung der semantischen Merkmale) und die Referenzidentität eine bedeutende Rolle. Diese Wiederaufnahmen können mit den Themahinweisen im Text zusammenfallen.

Die zusammenhängenden Wörter auf der Ebene der Kohärenz bilden die Isotopien. Diese beruhen auf der semantischen Äquivalenz zwischen mindestens zwei sprachlichen Einheiten. Die Isotopiebeziehungen werden bei der Teilthemauntersuchung im Gesamttext ergänzend angewendet, um die konkret zusammengehörenden Teil-themahinweise im Text und ihre Zugehörigkeit zu einem Referenzbereich zu erkennen.

Auch die Wortbildung hat Anteil an der Textgestaltung. Die Wortbildungs-produkte können im Text oft ohne Schwierigkeiten entdeckt werden, deswegen eignen sie sich als einleitender Schritt bei der Suche nach den relevanten Belegen. Bei der Be-wertung, ob sie sich als Themahinweise auf den gleichen Referenzbereich beziehen, wird der Kontext in Betracht gezogen. Diese Mittel dienen dazu, die relevanten Wortschatzelemente und ihre Beziehungen im Text als Themahinweis zu identifizieren.

Die Betrachtungen im fünften Kapitel konzentrieren sich auf das Auftreten des Teilthemas im Verlauf des Gesamttextes. Das Verständnis des Textthemas in dieser Arbeit geht davon aus, dass das Thema eines Textes besagt, wovon der Text handelt. Es stellt den Kern des Textinhalts dar. Innerhalb des Textthemas können mehrere Teilthemen auftreten, die sich z. B. in einzelnen Textteilen realisieren. Die Teilthemen spezifizieren das Hauptthema, sie sind ihm untergeordnet und können durch Zusam-

menfassungen oder Schlüsselwörter, die die betreffenden Textabschnitte vertreten, erkannt werden. Eine weitere Spezifizierung der Teilthemen erfolgt durch Subthemen.

In dem Roman von Günter Grass werden zwei Teilthemen abgegrenzt: ‚Liebesgeschichte' und ‚Friedhofsgesellschaft'. Diese Teilthemen werden durch die Beobachtung des Gesamttextinhalts und durch die Angaben in der Frequenz-Wortformenliste erkannt. Auch die Teilthemen können weiter in untergeordnete Subthemen gegliedert und spezifiziert werden. Ein Beispiel für ein Subthema ist der ‚Friedhof', der von der Friedhofsgesellschaft betrieben wird. Im Text kommen konkrete thematische Hinweise vor, die den Textrezipienten den Zusammenhang zu dem jeweiligen Referenzbereich signalisieren, welche anzeigen, worum es im Mitgeteilten geht. Im Laufe des Textes werden die einzelnen Phasen der Teilthemaprogression: die Einführung, Beibehaltung, Entwicklung und der Abschluss eines Teilthemas beobachtet.

Die Themahinweise decken sich oft mit den Kohäsions- und Kohärenzmitteln, daher sind diese Hinweise besser auffindbar; z. B. deckt sich die Kohäsion mit der Repetition identischer Lexeme mit dem Themabeibehaltungshinweis durch Rekurrenz. Im Bereich der Kohärenz wird ein weiteres Beispiel angeführt, in dem sich die semantische Progression mit den semantischen Themaentwicklungshinweisen deckt: […] *wird ihm die von der Witwe vorgeschlagene »Polnisch-Deutsch-Litauische Friedhofsgesellschaft« wichtig.* […] *Die litauische Komponente wertet er als »einleuchtend und wünschenswert«,* [...] (GGU, S. 53) Die litauische Komponente ist ein Teil der Polnisch-Deutsch-Litauischen Friedhofsgesellschaft. Wird dieser Fall mit den Kohärenzmitteln beschrieben, dann liegt die Beziehung der Isotopie vor. Aus dem Kontext wird die semantische Beziehung zwischen der Friedhofsgesellschaft und ihrer Komponente sichtbar; oder eben der semantische Themaentwicklungshinweis, denn die Komponente referiert auf den Referenzbereich ‚Friedhofsgesellschaft'.

Im anschließenden Abschnitt wird eine Probeanalyse durchgeführt, ihr Ziel ist es, die Vorgehensweise bei der Untersuchung im Gesamttext zu entwerfen. Die Analyse des ausgewählten Referenzbereichs ‚Engel' wird in vier Schritte aufgeteilt. Zuerst wird nach dem Vorkommen identischer Lexeme für das Wort ‚Engel' gesucht, um die Grenze des untersuchten Textabschnitts festzustellen. Die Menge der Belege wird mit anderen Lexemen ergänzt, die einen Bezug auf ‚Engel' haben (wie Synonyme oder andere referenziell identische Ausdrücke). Die textuelle Zugehörigkeit der Belege zueinander

wird durch die Kohäsions- und Kohärenzbeziehungen veranschaulicht. Abschließend werden die einzelnen Vorkommensfälle als Hinweise auf das Teilthema beschrieben.

Nicht alle diese Schritte wurden für die Untersuchung im Gesamttext erneut durchgeführt. Einige dieser Schritte wurden in vorherigen Abschnitten der Arbeit bewältigt, wie die Analyse der Frequenz-Wortformenliste, die das Teilthema ‚Friedhofs-gesellschaft' als den für die Untersuchung relevanten Referenzbereich bestimmte.

Dieses Kapitel wird mit der Untersuchung des ausgewählten Teilthemas abgeschlossen. Die Beziehungen zwischen den Wörtern im Text werden oft nur durch den Kontext sichtbar, deswegen wurde am Anfang der Inhalt des behandelten Romans vorgestellt. Der Inhalt zeigt, dass das Teilthema ‚Friedhofsgesellschaft' über den ganzen Text verteilt ist. Bei dem Referenzbereich ‚Engel' wurde als einleitender Schritt der Analyse die Suche nach identischen Lexemen durchgeführt, um die grobe Grenze des zu untersuchenden Textabschnitts festzulegen. Da die Friedhofsgesellschaft im ganzen Textverlauf des Romans vorkommt, wurde dieser Schritt hier nicht mehr durchgeführt. Das Sammeln der relevanten Belege wird so durchgeführt, dass sie gleich beim Lesen des Textes exzerpiert werden. Weitere zugehörende Belege werden in Bezug auf ihre Zusammengehörigkeit zueinander im Rahmen der erwähnten Wortschatzgruppierung bewertet. Die Verwendung der Prinzipien der Kohäsions- und Kohärenzbeziehungen unterstützt die Identifikation der Beziehung der aufzunehmenden Belege. Während der Arbeit mit Gesamttext wird nach den konkreten Hinweisen, die die Einführung, Beibehaltung, Entwicklung und den Abschluss des Teilthemas signalisieren, gesucht.

Die Themahinweise sind durch konkrete textuelle Einheiten vertreten. Trotzdem ist in manchen Fällen die Unterscheidung zwischen den jeweiligen Hinweistypen nicht ganz eindeutig. Ihre Differenzierung hängt nämlich oft davon ab, wie die Hinweise – jeweils mit Berücksichtigung des Kontextes – interpretiert werden. Während der Untersuchung wurde fast ausschließlich von Substantiven ausgegangen. In Bezug auf andere Wortarten wäre eine weitere Untersuchung ebenfalls empfehlenswert.

Bei der Analyse bieten sich auch andere Möglichkeiten, bestimmte Referenz-bereiche zu unterscheiden, wie z. B. die Themaentwicklung im Bereich der Personen-, Zeit-, Raum- oder Ereignisreferenz. Eine ausführlichere Analyse der einzelnen Referenzbereiche wurde hier jedoch nicht durchgeführt, da eine komplette Textanalyse der angeführten Aspekte nicht zum Ziel dieser Arbeit gehört.

Bibliographie

Textquellen

Brussig, Thomas: *Am kürzeren Ende der Sonnenallee*. Frankfurt am Main 2001.

Brussig, Thomas: *Na kratším konci ulice*. Übersetzt von Zoubková, Jana. Praha 2001.

Grass, Günter: *Žabí lamento.* Übersetzt von Karlach, Hanuš. Brno 1996.

Grass, Günter: *Unkenrufe*. Göttingen 1992.

Kratochvil, Jiří: *Nesmrtelný příběh aneb Život Soni Trocké-Sammlerové čili Román karneval.* Brno 2005.

Kratochvil, Jiří: *Unsterbliche Geschichte oder Das Leben der Sonja Trotzkij-Sammler oder Karneval.* Übersetzt von Liedtke, Kathrin und Vagadyová, Milka. Zürich 2000.

Wissenschaftliche Literatur

Agricola, Erhard: *Semantische Relationen im Text und im System.* Halle 1975.

Barz, Irmhild [u. a.]: *Wortbildung – praktisch und integrativ. Ein Arbeitsbuch*. Frankfurt am Main ^{3}2004.

Beaugrande, Robert-Alain de / Dressler, Wolfgang Ulrich: *Einführung in die Textlinguistik.* (= Konzepte der Sprach- und Literaturwissenschaft 28), Tübingen 1981.

Biber, Douglas [u. a.]: *Corpus linguistics. Investigating Language Structure and Use*. Cambridge ^{5}2006.

Brinker, Klaus / Hagemann, Jörg: *Themenstruktur und Themenentfaltung in Gesprächen.* In: Brinker, Klaus [u. a.] (Hg.): Text- und Gesprächslinguistik. Berlin – New York 2001. S. 1252–1263.

Brinker, Klaus: *Linguistische Textanalyse. Eine Einführung in Grundbegriffe und Methoden.* Berlin ^{6}2005.

Brunner, Gerhard / Ecker Bernhard: *CorelDRAW 10. Professionelle Grafik für Druck und Internet.* Böblingen 2001.

Dias, Idalete: *Das deutsch-portugiesische PORTDE-Korpus*. In: Schwitalla, Johannes / Wegstein, Werner (Hg.): *Korpuslinguistik deutsch: synchron – diachron – kontrastiv. Würzburger Kolloquium 2003*. Tübingen 2005, S. 255–258.

Dubinin, Sergej / Vadayev, Sergej / Smolskaja, Julia: *Probleme und Schwerpunkte eines modernen russisch-deutschen Textkorpus*. In: Schwitalla, Johannes / Wegstein, Werner (Hg.): *Korpuslinguistik deutsch: synchron – diachron – kontrastiv. Würzburger Kolloquium 2003*. Tübingen 2005, S. 259–266.

Erben, Johannes: *Deutsche Grammatik*. Frankfurt am Main 1975.

Fleischer, Wolfgang / Barz, Irmhild: *Wortbildung der deutschen Gegenwartssprache*. Tübingen 1995.

Friedl, Jeffrey E. F.: *Reguläre Ausdrücke*. Deutsche Übersetzung: Karrer, Andreas. Köln [u. a.] ^{2}2003.

Fritz, Thomas A.: *Der Text*. In: Duden. Die Grammatik. Mannheim [u. a.] ^{7}2005, S. 1067–1174.

Hausendorf, Heiko / Kesselheim, Wolfgang: *Textlinguistik fürs Examen*. Göttingen 2008.

Greimas, Algirdas Julien: *Die Isotopie der Rede*. In: Kallmeyer, Werner [u. a.]: *Lektürekolleg zur Textlinguistik 2*. Frankfurt am Main 1974, S. 126–152.

Hackl-Rößler, Sabine: *Textstruktur und Textdesign*. Tübingen 2006.

Kenney, Anne R.: *Moving theory into practice*. Mountain View 2000.

Klein, Wolfgang / von Stutterheim, Christiane: *Quaestio und referentielle Bewegung in Erzählungen*. Linguistische Berichte 109, 1987, S. 163–183.

Kosek, Jiří: *XML pro každého, podrobný průvodce*. Praha 2000.

Kotůlková, Veronika: *Deutsche Determinativkomposita und ihre Äquivalente im Tschechischen. Eine korpusbasierte kontrastive Studie*. Saarbrücken 2009.

Kratochvílová, Iva: *Analysen in Spezialkorpora: Die würde-Konstruktion in narrativen Texten.* In: Kratochvílová, Iva / Wolf, Norbert Richard: *Kompendium Korpuslinguistik. Eine Bestandaufnahme aus deutsch-tschechischer Perspektive.* Heidelberg 2010, S. 171–177.

Kraus, Helmut: *Scans, Prints & Proofs. Bessere Ergebnisse beim Scannen und Drucken*. Bonn 2001.

Kraus, Helmut: *Scannen: mit Desktopscannern zum perfekten Bild*. Bonn ^{2}1998.

Lemnitzer, Lothar / Zinsmeister, Heike: *Korpuslinguistik. Eine Einführung*. Tübingen 2006.

Linke, Angelika / Nussbaumer, Markus / Portmann, Paul R.: *Studienbuch Linguistik*. Tübingen 2004.

Lutzeier, Peter Rolf: *Lexikologie: ein Arbeitsbuch*. Tübingen 1995.

Maschke, Thomas: *Digitale Bildbearbeitung*. *Bildbearbeitung, Farbmanagement, Bildausgabe.* Berlin / Heildelberg [u. a.] 2004.

McEnery, Tony: *Corpus linguistics*. In: Mitkov, Ruslan (Hg.): *Computational linguistics*. Oxford / New York [u. a.] 2003.

McEnery, Tony / Xiao Richard / Tono Yukio: *Corpus-Based Language Studies, an advanced resource book.* London 2006.

McEnery, Tony / Wilson, Andrew: *Corpus Linguistics, An Introduction.* Edinburgh ^{2}2001.

Nyman, Mattias: *4 Farben 1 Bild*. Berlin ^{3}1999.

O'Keeffe, Anne / McCarthy Michael: *The Routledge Handbook of Corpus Linguistics*. London / New York 2010.

Osterberg, Jürgen. *GIMP 2. Anspruchsvolle Bildbearbeitung unter Linux, Windows und Mac OS X*. Heidelberg ^{2}2005.

Rastier, François: *Systematik der Isotopien.* In: Kallmeyer, Werner [u. a.] *Lektürekolleg zur Textlinguistik 2.* Frankfurt am Main 1974, S. 153–190.

Ray, Erik T.: *Einführung in XML*. Beijing [u. a.] 2001.

Römer, Christine / Matzke, Brigitte: *Lexikologie des Deutschen. Eine Einführung*. Tübingen ^{2}2005.

Scherer, Carmen: *Korpuslinguistik*. Heidelberg 2006.

Schippan, Thea: *Lexikologie der deutschen Gegenwartssprache*. Tübingen 1992.

Schlicht, Hans-Jürgen: *Digitale Bildverarbeitung mit dem PC. Scanner Drucker Video Multimedia.* Bonn 1993.

Thiel, Gisela: *Isotopie. Eine textlinguistische Kategorie im Dienst der Übersetzung*. In: Lauer, Angelika (Hg.): *Übersetzungswissenschaft im Umbruch: Festschrift für Wolfram Wilss zum 70. Geburtstag.* Tübingen 1996, S. 59–68.

Tidwell, Dough: *XSLT*. Deutsche Übersetzung: Lichtenberg, Kathrin und Brodacki, Olaf. Köln [u. a.] 2002

Trier, Jost: *Der deutsche Wortschatz im Sinnbezirk des Verstandes. Die Geschichte eines sprachlichen Feldes.* Band 1 (Von den Anfängen bis zum Beginn des 13. Jahrhunderts) Heidelberg ^{2}1973.

Waldraff, Thomas: *Digitale Bildauflösung. Grundlagen, Auflösungsbestimmung, Anwendungsbeispiele*. Berlin / Heidelberg 2004.

Wolf, Norbert Richard: *Korpora in der Korpuslinguistik*. In: Kratochvílová, Iva / Wolf, Norbert Richard: *Kompendium Korpuslinguistik. Eine Bestandaufnahme aus deutsch-tschechischer Perspektive.* Heidelberg 2010, S. 17–25.

Wolf, Norbert Richard: *Textsyntax und / oder Textstilistik*. In: Fritz, Thomas, A. (Hg.): *Literaturstil – sprachwissenschaftlich.* Heidelberg 2008, S. 57–69.

Wolf, Norbert Richard: *Wörter bilden. Grundzüge der Wortbildungslehre*. In: Dittmann, Jürgen / Schmidt, Claudia (Hg.): *Über Wörter – Grundkurs Linguistik*. Freiburg im Breisgau / Rombach 2002. S. 59–86.

Elektronische Quellen

AntConc. *Startseite* [online]. Tokyo, 2010 [zit.: 6. 2. 2010]. Zugänglich unter WWW: <http://www.antlab.sci.waseda.ac.jp/antconc_index.html>

Blümm, Mirjam: *Leitlinien zur Kodierung des Campe-Wörterbuchs in TEI P5* [online]. Würzburg, Version Januar 2009 [zit.: 20. 2. 2010]. Zugänglich unter WWW: <http://www.textgrid.de/fileadmin/TextGrid/veroeffentlichungen/Leitlinien-zur-Kodierung-des-Campe-Woerterbuchs-in-TEI-P5.pdf>

Corpus Encoding Standard. *Startseite* [online]. Version 1.5, 2000 [zit.: 29. 12. 2010]. Zugänglich unter WWW: <http://www.cs.vassar.edu/CES>

Das deutsch-tschechische Parallelkorpus: DeuCze. *Startseite* [online]. Würzburg [zit.: 23. 6. 2010]. Zugänglich unter WWW: <deucze.org | http://www.deucze.germanistik.uni-wuerzburg.de>

Das Institut für Deutsche Sprache. *Cosmas II* [online]. Mannheim [zit.: 7. 5. 2010]. Zugänglich unter WWW: <https://cosmas2.ids-mannheim.de/cosmas2-web>

Das Institut für Deutsche Sprache. *Startseite* [online]. Mannheim [zit.: 7. 5. 2010]. Zugänglich unter WWW: <http://www.ids-mannheim.de>

Deutsche Forschungsgemeinschaft. *Wissenschaftliche Literaturversorgungs- und informationssysteme (LIS): DFG-Praxisregeln „Digitalisierung“* [online] Bonn, April 2009 [zit.: 24. 6. 2009]. Zugänglich unter WWW: <http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/download/praxisregeln_digitalisierung.pdf>

European Parliament Proceedings Parallel Corpus 1996-2009 [online]. 2000 [zit.: 18. 11. 2010]. Zugänglich unter WWW: <http://www.statmt.org/europarl>

Expert Advisory Group on Language Engineering Standards. *Homepage* [online]. [zit.: 29. 12. 2010]. Zugänglich unter WWW: <http://www.ilc.cnr.it/EAGLES/home.html>

Filozofická fakulta Univerzity Karlovy. *Český národní korpus* [online] Praha. [zit.: 7. 5. 2010]. Zugänglich unter WWW: <http://ucnk.ff.cuni.cz>

Francis, W. N. / Kucera, H.: *Brown Corpus Manual*. Brown University, Providence, Revised and Amplified1979 [zit.: 23. 11. 2010]. Zugänglich unter WWW: <http://icame.uib.no/brown/bcm.html>

Koehn, Philipp: *Europarl: A Parallel Corpus for Statistical Machine Translation* [online]. University of Edinburgh, 2005 [zit.: 18. 11. 2010]. Zugänglich unter WWW: <http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/europarl-mtsummit05.pdf>

The Department of Computational Linguistics and Phonetics in Saarbrücken [u. a.]. *TIGER PROJECT. Linguistic Interpretation of a German Corpus*. Saarbrücken / Stuttgart / Potsdam, 2007 [zit.: 18. 11. 2010]. Zugänglich unter WWW: <http://www.ims.uni-stuttgart.de/projekte/TIGER>

TEI: Text Encoding Initiative. *Startseite* [online]. 2007 [zit.: 13. 7. 2009]. Zugänglich unter WWW: <http://www.tei-c.org/index.xml>

The TEI Consortium. *TEI P5: Guidelines for Electronic Text Encoding and Interchange* [online]. Oxford / Providence / Charlottesville / Nancy, Version 1.4.1. Juli 2009 [zit.:

13. 7. 2009]. Zugänglich unter WWW: <http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf>

Vavřín, Martin / Rosen, Alexandr: *Korpus InterCorp* [online]. Praha. [zit.: 7. 5. 2010]. Zugänglich unter WWW: <http://korpus.cz/intercorp-info.php>

Wikipedia. Die freie Enzyklopädie. *Startseite* [online]. San Francisco: Wikimedia Foundation Inc., 2001- [25. 1. 2011]. Zugänglich unter WWW: <http://de.wikipedia.org>

Wegstein, Werner / Blümm, Mirjam / Seipel, Dietmar / Schneiker, Christian: *Digitalisierung von Primärquellen für die TextGrid-Umgebung: Modellfall Campe-Wörterbuch* [online]. Würzburg, Version Oktober 2009 [zit.: 20. 2. 2010]. Zugänglich unter WWW: <http://www.textgrid.de/fileadmin/TextGrid/reports/TextGrid_R4_1.pdf>

Lexika

Auberle, Anette (Red.): *Duden. Herkunftswörterbuch: Etymologie der deutschen Sprache.* Mannheim / Leipzig / Wien / Zürich [3]2001.

Dornseiff, Franz: *Der deutsche Wortschatz nach Sachgruppen.* Berlin / New York [7]1970.

Duden – Deutsches Universalwörterbuch, 6. Aufl. Mannheim 2006 [CD-ROM].

Duden – Das Synonymwörterbuch, 4. Aufl. Mannheim 2007 [CD-ROM].

Glück, Helmut (Hg.): Metzler Lexikon Sprache. Stuttgart / Weimar 2000.

Scholze-Stubenrecht, Werner: *Duden, die deutsche Rechtschreibung.* Mannheim 2001.

Wehrle, Hugo / Eggers, Hans: *Deutscher Wortschatz: ein Wegweiser zum treffenden Ausdruck.* 1. Aufl., 17. Nachdr. Stuttgart 1993.

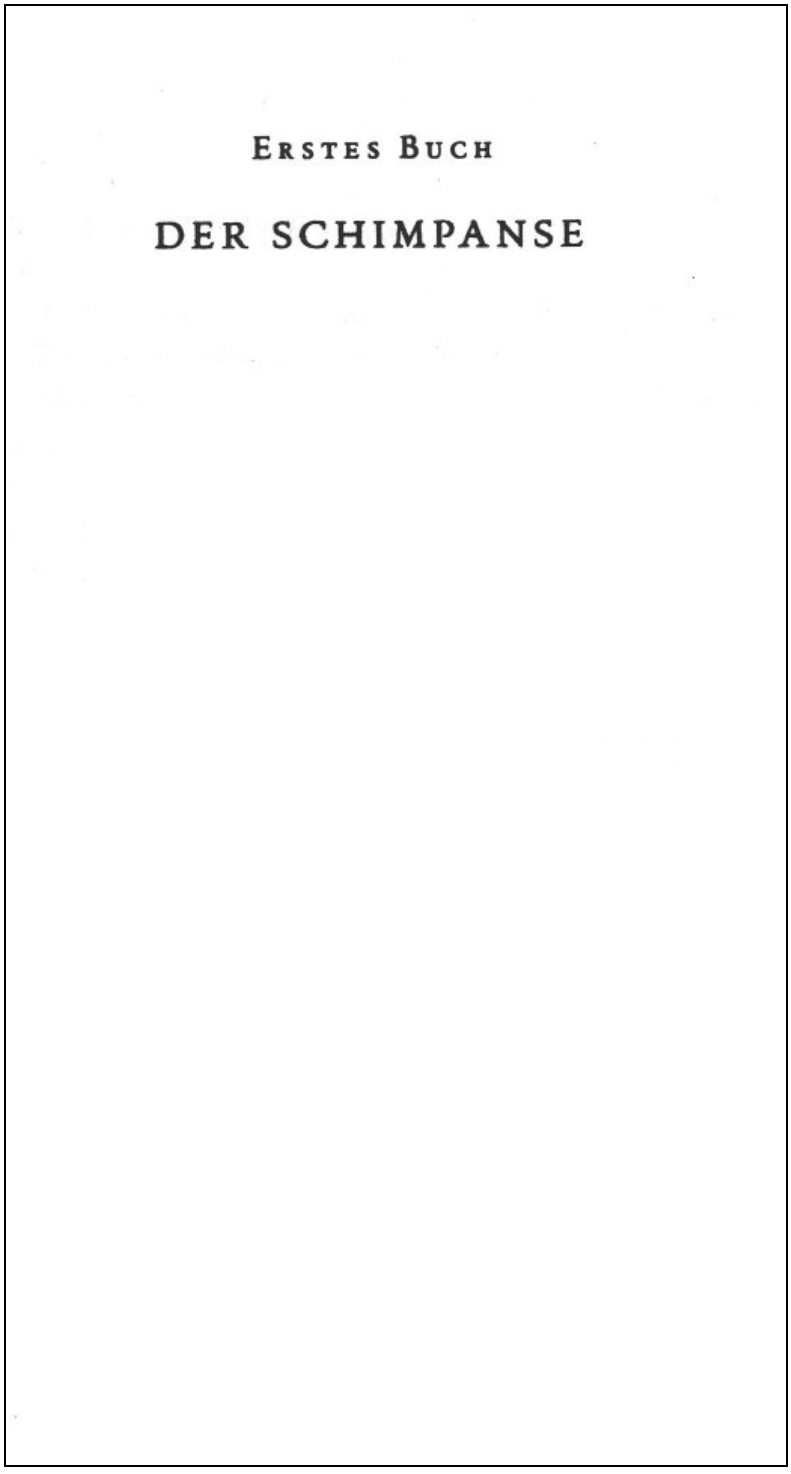
Wahrig, Gerhard: *Deutsches Wörterbuch.* Gütersloh / München [7]2002.

Anhang

Musterseiten

Kratochvil, Jiří: *Unsterbliche Geschichte oder Das Leben der Sonja Trotzkij-Sammler oder Karneval.* Übersetzt von Liedtke, Kathrin und Vagadyová, Milka. Zürich 2000.

a) Überschrift des Buches als Teil des Romans (S. 7)



b) Anfang des ersten Kapitels (S. 9)

1

Die Stimme meines Herrn

Geboren wurde ich, wenn Sie das wirklich hören wollen, meine Herren, geboren wurde ich in der Nacht vom 31. Dezember 1899 auf den 1. Januar 1900, und mein Vater war der Sohn eines orthodoxen russischen Priesters und meine Mutter eine Deutsche (ihren Eltern – und später ihrem Bruder – gehörte ein großes Gut in Landskron am Fuße des Adlergebirges). Ich wurde zu Hause geboren, damals machte man das noch so. Meine Hebamme war Magda, eine Ungarin aus Preßburg. Meine Geburtsstadt Brünn. Die genaue Stelle dann ein Bett, eine Pritsche im dritten Stock eines großen Mietshauses in der Ferdinandstraße (später Masarykstraße, noch später Hermann-Göring-Straße und noch später wieder Masarykstraße und noch später Straße des Sieges und dann kurz wieder Masarykstraße und dann wieder lange Zeit Straße des Sieges und heute wieder Masarykstraße).

Jede Geburt ist, falls Sie das nicht wissen sollten, meine Lieben, für alle Beteiligten auch ein Zirkus von Emotionen. Die Hebamme hielt mich und schrie etwas auf ungarisch und slowakisch. Mutter versuchte, mich anzusehen, und obwohl man sie daran zu hindern versuchte und mit aller Gewalt ins Bett zurückdrückte, stützte sie sich auf die Ellenbogen, richtete sich auf wie eine Kobra und murmelte einen Augenblick hastig etwas auf tschechisch und gleich darauf wieder auf deutsch. Vater kniete auf dem Boden, und er, der es mit dem Glauben bisher nicht so genau genommen hatte, betete jetzt auf russisch, und in diesen großen orthodoxen Strom, der sich aus ihm herauswälzte, flossen auch

9

c) Gedicht (S. 172)

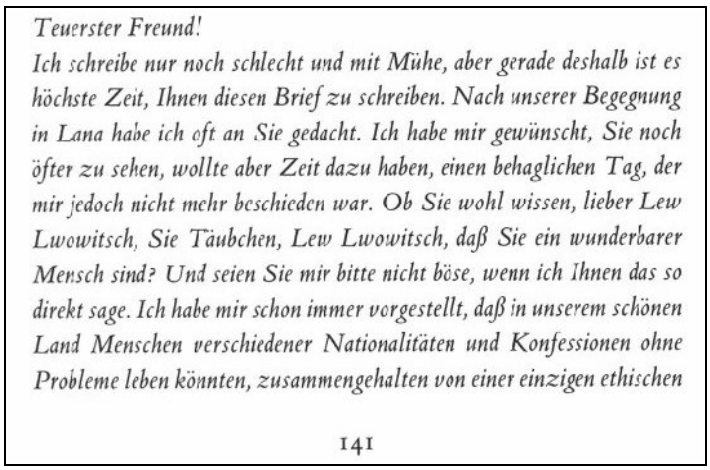
Der ewige Aufschub der Aktion wirkte sich zersetzend bis geradezu vernichtend auf die Moral meiner Wolfssoldaten aus (haben Sie's gemerkt, ich sage bereits: *meiner*), und so reichte es nicht aus, mich ihnen – um sie überhaupt über eine so lange Zeit hinweg in der erforderlichen Kondition zu halten – lediglich als Befehlshaberin und Übungsleiterin zu widmen, denn in jedem dieser zotteligen Wolfsfelle steckte ein sensibler und, wie mir manchmal schien, übersensibler Iwan (Aljoscha, Boris, Fjodor, Sergej), und in jedem von ihnen verbarg sich wiederum ein empfindsames bis überempfindsames menschliches Herz, das auch seine Nahrung brauchte (und nicht zu vergessen die unsterbliche russische Seele, die besonders schwer zu sättigen war).

Die ganze Nacht über trainierten sie hart, bis ihnen der Schweiß herunterrann und in den Pfützen unter den Bäumen zum stärksten Slibowitz (oder, wenn Sie wollen, Wodka) destillierte, und am Morgen sprangen sie dann in die fünf Baumkronen, die den sechsten Baum umringten, unter dem ich mich niedergelassen hatte. Und ich appelliere abermals an Ihr Vorstellungsvermögen: fünf Bäume also und ein sechster in der Mitte, von dem aus ich ihnen zum Einschlafen das Märchen von Ruslan und Ludmila erzählte, jawohl, dasselbe, das mir einst Vater als Gute-Nacht-Geschichte erzählt hatte und das immer, hören Sie, wie folgt begann:

U lukomorja dub zeljonyj
zlataja cepj na dubje tom:
i dnjom i nočju kot učonyj
vsjo chodit po cepi krugom …

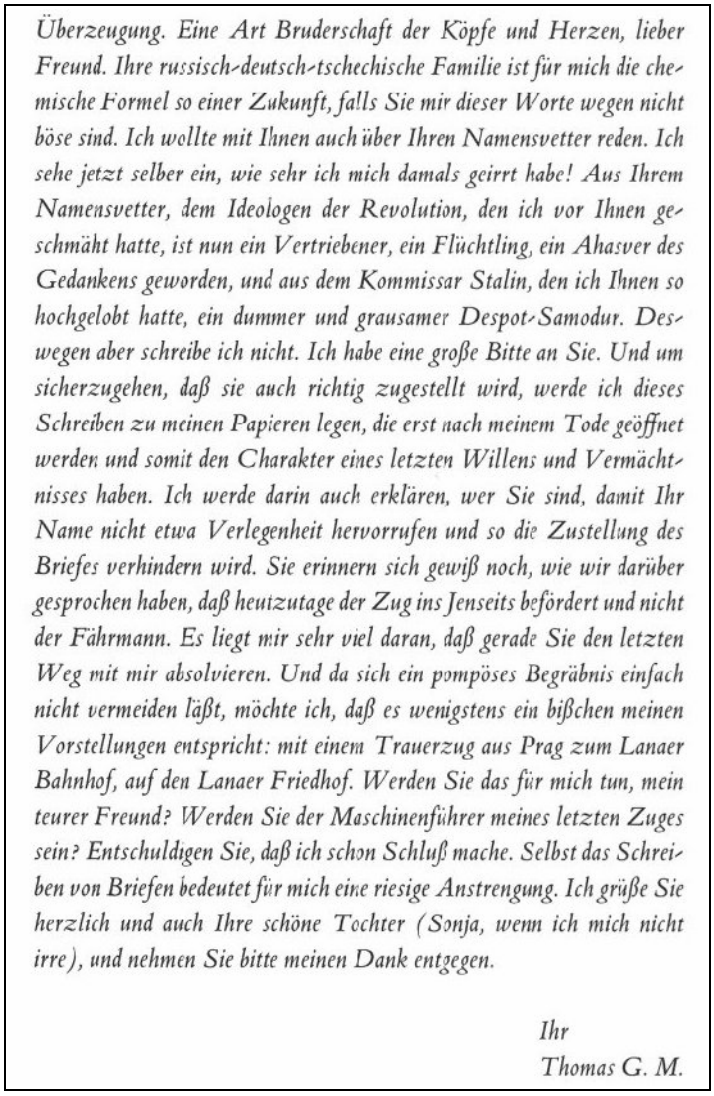
Und wie ich so erzählte, hörte ich, wie die Werwölfe auf den Bäumen um mich herum nacheinander einschliefen, in Schlaf versanken *(immer im Kreis herum läuft der gelehrte Kater um die grüne Eiche, und die goldene Kette läutet),* bis ich allein blieb, mutter-

d) Brief: (S. 141 Teil 1 und S. 142 Teil 2)



Teuerster Freund!
Ich schreibe nur noch schlecht und mit Mühe, aber gerade deshalb ist es höchste Zeit, Ihnen diesen Brief zu schreiben. Nach unserer Begegnung in Lana habe ich oft an Sie gedacht. Ich habe mir gewünscht, Sie noch öfter zu sehen, wollte aber Zeit dazu haben, einen behaglichen Tag, der mir jedoch nicht mehr beschieden war. Ob Sie wohl wissen, lieber Lew Lwowitsch, Sie Täubchen, Lew Lwowitsch, daß Sie ein wunderbarer Mensch sind? Und seien Sie mir bitte nicht böse, wenn ich Ihnen das so direkt sage. Ich habe mir schon immer vorgestellt, daß in unserem schönen Land Menschen verschiedener Nationalitäten und Konfessionen ohne Probleme leben könnten, zusammengehalten von einer einzigen ethischen

141



Überzeugung. Eine Art Bruderschaft der Köpfe und Herzen, lieber Freund. Ihre russisch-deutsch-tschechische Familie ist für mich die chemische Formel so einer Zukunft, falls Sie mir dieser Worte wegen nicht böse sind. Ich wollte mit Ihnen auch über Ihren Namensvetter reden. Ich sehe jetzt selber ein, wie sehr ich mich damals geirrt habe! Aus Ihrem Namensvetter, dem Ideologen der Revolution, den ich vor Ihnen geschmäht hatte, ist nun ein Vertriebener, ein Flüchtling, ein Ahasver des Gedankens geworden, und aus dem Kommissar Stalin, den ich Ihnen so hochgelobt hatte, ein dummer und grausamer Despot-Samodur. Deswegen aber schreibe ich nicht. Ich habe eine große Bitte an Sie. Und um sicherzugehen, daß sie auch richtig zugestellt wird, werde ich dieses Schreiben zu meinen Papieren legen, die erst nach meinem Tode geöffnet werden und somit den Charakter eines letzten Willens und Vermächtnisses haben. Ich werde darin auch erklären, wer Sie sind, damit Ihr Name nicht etwa Verlegenheit hervorrufen und so die Zustellung des Briefes verhindern wird. Sie erinnern sich gewiß noch, wie wir darüber gesprochen haben, daß heutzutage der Zug ins Jenseits befördert und nicht der Fährmann. Es liegt mir sehr viel daran, daß gerade Sie den letzten Weg mit mir absolvieren. Und da sich ein pompöses Begräbnis einfach nicht vermeiden läßt, möchte ich, daß es wenigstens ein bißchen meinen Vorstellungen entspricht: mit einem Trauerzug aus Prag zum Lanaer Bahnhof, auf den Lanaer Friedhof. Werden Sie das für mich tun, mein teurer Freund? Werden Sie der Maschinenführer meines letzten Zuges sein? Entschuldigen Sie, daß ich schon Schluß mache. Selbst das Schreiben von Briefen bedeutet für mich eine riesige Anstrengung. Ich grüße Sie herzlich und auch Ihre schöne Tochter (Sonja, wenn ich mich nicht irre), und nehmen Sie bitte meinen Dank entgegen.

Ihr
Thomas G. M.

Buchstabenmuster und Kodierungsvorgaben für die Texterfassung

Stellvertretend wurde für die Beschreibung dieser Korpustext gewählt: Kratochvil, Jiří: *Unsterbliche Geschichte oder Das Leben der Sonja Trotzkij-Sammler oder Karneval.* Übersetzt von Liedtke, Kathrin und Vagadyová, Milka. Zürich 2000.

Die angeführten Zeichen sind als Ausgangsbasis zu betrachten, bei der Bearbeitung des Benutzermusters muss auf die Druck- und Scanqualität der jeweiligen Stellen geachtet werden.

Verwendeter Zeichensatz: Unicode UTF-8

Die Mustertranskription gilt jeweils nur für den unterstrichenen Buchstaben.

Normalschrift		Kapitälchenschrift		Kursivschrift	
Groß-buchstabe	Klein-buchstabe	Groß-buchstabe	Klein-buchstabe	Groß-buchstabe	Klein-buchstabe
Aufgabe	auf	AUS	DAS	*Anna*	*aber*
A (Aufgabe)	a (auf)	A (Aus)	a (Das)	A (Anna)	a (aber)
- - - -	Jirásekviertel	- - - -	- - - -	- - - -	*právo*
	á (Jirásekviertel)				á (právo)
Ärmsten	hätte	ÄGÄISCHEN	LÄUFER	- - - -	*Oberfläche*
Ä (Ärmsten)	ä (hätte)	Ä (Ägäischen)	ä (Läufer)		ä (Oberfläche)
Bruno	böse	BLUTIGE	GEBOREN	*Braut*	*habe*
B (Bruno)	b (böse)	B (Blutige)	b (geboren)	B (Braut)	b (habe)
Celebes	dich	CHARON	ICH	*Cervus*	*Ich*
C (Celebes)	c (dich)	C (Charon)	c (Ich)	C (Cervus)	c (Ich)
- - - -	Horáček	- - - -	- - - -	- - - -	*spáči!*
	č (Horáček)				č (spáči)
Dinge	die	DIE	BEERDIGEN	*Dorf,*	*deshalb*
D (Dinge)	d (die)	D (Die)	d (beerdigen)	D (Dorf)	d (deshalb)
Ei	ein	EDISON	LEBE	*Erzengel*	*es*
E (Ei)	e (ein)	E (Edison)	e (lebe)	E (Erzengel)	e (es)
Époque	Straßencafés	- - - -	- - - -	- - - -	*Rudé*
É (Époque)	é (Straßencafés)				é (Rudé)

- - - -	Alžbětka	- - - -	ALŽBĚTKA	- - - -	- - - -
	ě (Alžbětka)		ě (Alžbětka)		

Fachmann	triftige	FRAU	DARAUF	*Ferkel*	*öfter*
F (Fachmann)	f (triftige)	F (Frau)	f (darauf)	F (Ferkel)	f (öfter)

- - - -	betrifft	- - - -	- - - -	- - - -	*geöffnet*
	[Ligatur] ff (betrifft)				[Ligatur] ff (geöffnet)

- - - -	zufiel	- - - -	- - - -	- - - -	*empfing*
	[Ligatur] fi (zufiel)				[Ligatur] fi (empfing)

- - - -	pflegen	- - - -	- - - -	- - - -	*pflügen*
	[Ligatur] fl (pflegen)				[Ligatur] fl (pflügen)

Gründe	frage	GLÜHEN	GLEICH	*Georgij*	*gerade*
G (Gründe)	g (frage)	G (Glühen)	g (gleich)	G (Georgij)	g (gerade)

Hallodri	heißt	HERRN	ZIEHT	*Häuschen,*	*Nach*
H (Hallodri)	h (heißt)	H (Herrn)	h (zieht)	H (Häuschen)	h (Nach)

Iljitsch	die	IM	IN	*Ikonen*	*in*
I (Iljitsch)	i (die)	I (Im)	i (im)	I (Ikonen)	i (in)

- - - -	Lubomír	- - - -	- - - -	- - - -	- - - -
	í (Lubomír)				

Jedes	jetzt	JAHRE	SARAJEVO	*Jahre*	*jedoch*
J (Jedes)	j (jetzt)	J (Jahre)	j (Sarajevo)	J (Jahre)	j (jedoch)

Karpfen	keine	KLUGHEIT	HINKENDE	*Karenina,*	*direkt*
K (Karpfen)	k (keine)	K (Klugheit)	k (hinkende)	K (Karenina)	k (direkt)

Laufe	lieb	LANDSKRON	TEUFEL	*Lazarus*	*behaglichen*
L (Laufe)	l (lieb)	L (Landskron)	l (Teufel)	L (Lazarus)	l (behaglichen)

Mlock	mir	MEER	MEINES	*Menschen*	*mit*
M (Mlock)	m (mir)	M (Meer)	m (meines)	M (Mensch)	m (mit)

Nikolaus	nur	- - - -	DEN	*Nikolaus*	*nur*
N (Nikolaus)	n (nur)		n (den)	N (Nikolaus)	n (nur)

- - - -	- - - -	- - - -	- - - -	- - - -	*Soňo,*
					ň (Soňo)

Onkel	so	OBERST	GEHORSAM	*Oberfläche*	*noch*
O (Onkel)	o (so)	O (Oberst)	o (gehorsam)	O (Oberfläche)	o (noch)

Öffne	Gehröcken	- - - -	SCHÖNE	- - - -	*höchste*
Ö (Öffne)	ö (Gehröcken)		ö (Schöne)		ö (höchste)

Popen	passierte	PALÄSTEN	SPINOZA	*Petersburg*	*Mop*
P (Popen)	p (passierte)	P (Palästen)	p (Spinoza)	P (Petersburg)	p (Mop)

Quere	Aquarium	- - - -	- - - -	*Qualität*	*Jean-Jacques*
Q (Quere)	q (Aquarium)			Q (Qualität)	q (Jean-Jacques)

Reste	redete	REDL	ERDBEBEN	*Regalecus*	*mir*
R (Reste)	r (redete)	R (Redl)	r (Erdbeben)	R (Ragalecus)	r (mir)

Řečko- witz	Pařízek	- - - -	- - - -	- - - -	- - - -
Ř[107] (Řečkowitz)	ř (Pařízek)				

Sie	sag	STIMME	DAS	*Sonntags*	*unserer*
S (Sie)	s (sag)	S (Stimme)	s (das)	S (Sonntags)	s (unserer)

- - - -	Luboš	- - - -	- - - -	- - - -	*ničenskaja*
	š (Luboš)				š (ničenskaja)

- -	mußte,	- -	- - - -	- -	*Großonkel*
	ß (mußte)				ß (Großonkel)

Tat	fortsetzten	TIERE	GARTEN	*Tiefe*	*Bythios*
T (Tat)	t (fortsetzen)	T (Tiere)	t (Garten)	T (Tiefe)	t (Bythios)

Und	und	UND	UNTERRICHTE	*Unruhen*	*und*
U (Und)	u (und)	U (Und)	u (unterrichte)	U (Unruhen)	u (und)

Überschrift	kümmerte	- - - -	RÜCKKEHR	*Überzeugung.*	*Flüchtling,*
Ü (Überschrift)	ü (kümmerte)		ü (Rückkehr)	Ü (Überzeugung)	ü (Flüchtling)

Vorsicht	verzauberte	VIERTES	VIVA	*Verkaufte*	*von*

107 Das Wort ‚Řečkowitz' steht im Text am Zeilenende getrennt.

V (Vorsicht)	v (verzauberte)	V (Viertes)	v (viva)	V (Verkaufte)	v (von)

Weißt	während	WANN	WILL	*Wurzeln*	*gewünscht,*
W (Weißt)	w (während)	W (Wann)	w (will)	W (Wurzeln)	w (gewünscht)

- - - -	Luxusvilla	- - - -	- - - -	- - - -	*orthodoxen*
	x (Luxusvilla)				x (orthodoxen)

Yperits	Märtyrer	- - - -	MYSTISCHE	*York*	*zeljonyj*
Y (Yperits)	y (Märtyrer)		y (mystische)	Y (York)	y (zeljonyj)

- - - -	Malinovský-Platz	- - - -	- - - -	- - - -	- - - -
	ý (Malinovský-Platz)				

Zeiten	zufiel	ZWÖLF	LAZARUS	*Zaren*	*zu*
Z (Zeiten)	z (zufiel)	Z (Zwölf)	z (Lazarus)	Z (Zaren)	z (zu)

Žyla	Alžbětkas	- - - -	ALŽBĚTKA	- - - -	*ničtožestvo*
Ž (Žyla)	ž (Alžbětkas)		ž (Alžbětka)		ž (ničtožestvo)

Buchstaben des tschechischen Alphabets, die im deutschen Text nicht vertreten sind: Ď, ď, Ó, ó, Ť, ť, Ú, ú, Ů und ů, bei der tschechischen Zusammenrückung ch gibt es keine Sondermaßnahmen.

Sonderfall:

Die römischen Zahlen sind im Originaltext als Kleinbuchstaben der Kapitälchenschrift gedruckt. Bei der OCR-Bearbeitung muss besonders darauf geachtet werden, dass das Zeichen ı nicht als kleines l oder die Ziffer 1 interpretiert wird. Die Zahlen werden mit Kleinbuchstaben kodiert und in der Benutzeroberfläche des Korpus mithilfe der Formatierungstags in Kapitälchenschrift angezeigt.[108]

Franz Joseph ı.,	Franz Joseph i.,
in der ıc	in der ic
Nikolaus ıı.	Nikolaus ii.
Philipp ıv.	Philipp iv.
Ludwigs xvı.	Ludwigs xvi.
Nikolaus ıı.[Kursive]	Nikolaus ii.

108 Im tschechischen Text von Kratochvil stehen die römischen Zahlen in Normalschrift und Großbuchstaben.

Interpunktion:

- Komma – [,]

so, daß	WIEN, WIEN	*noch, wie*
so, daß	Wien, Wien	noch, wie

- Punkt – [.]

hatte. Wir	- - - -	*wird. Sie*
hatte. Wir		wird. Sie

- öffnende runde Klammer – [(]

(sie	- - - -	*(Sonja,*
(sie		(Sonja

- schließende runde Klammer – [)]

angereist)	- - - -	*irre)*
angereist)		irre)

- Bindestrich – [-]

Guth-Jarkovsky	KARPATO-UKRAINE	*Despot-Samodur.*
Guth-Jarkovsky	Karpato-Ukraine	Despot-Samodur

- Gedankenstrich – [–]

Eltern – und	MOS – EIN	*tel – Friede*
Eltern – und	mos – ein	tel – Friede

- Trennstrich – [-]

Frauenarbeiten und -fertigkeiten	- - - -	*ge-*
Frauenarbeiten und -fertigkeiten		ge-

- Ausrufezeichen – [!]

bekamen!	LIEBE!	*Freund!*
bekamen!	Liebe!	Freund!

- Fragezeichen – [?]

geboren?	GEBOREN?	*sind?*
geboren?	geboren?	sind?

- Apostroph – [’]

wir’s	- - - -	*sull’ali*
wir’s		sull’ali

- Doppelpunkt – [:]

hervor:	- - - -	*entspricht: mit*
hervor:		entspricht: mit

- Semikolon – [;]

nehmen;	- - - -	- - - -
nehmen;		

- Auslassungszeichen – […]

der … Rekruten	- - - -	*Pfeil …*
der … Rekruten		Pfeil …

- Anführungszeichen – öffnendes [»]

»Blutwürstchen	- - - -	- - - -
»Blutwürstechen		

- Anführungszeichen – schließendes [«]

Lubošek«	- - - -	- - - -
Lubošek«		

Ziffern:

1922	26.	1883	1941	1945	196	17.	1918	1951	1960
1 (1922)	2 (26)	3 (1883)	4 (1941)	5 (1945)	6 (196)	7 (17)	8 (1918)	9 (1951)	0 (1960)

Sonstiges:

Länge der Zeilen entspricht dem Original (Zeilenumbruch durch Enter-Taste)

Silbentrennung am Zeilenende wird als Trennstrich beibehalten.

Die Schriftformatierung wird mit den entsprechenden XML-Elementen verzeichnet:

- Kursivschrift: `<hi rend="italic"> … </hi>`
- Kapitälchenschrift: `<hi rend="smallCaps"> … </hi>`
- Fettschrift: `<hi rend="bold"> … </hi>` (im deutschen Kratochvil-Text nicht vertreten)

Bestimmung der Satzgrenze

Als Satzgrenze können Punkt, Frage- oder Ausrufezeichen, Auslassungszeichen vorkommen.

Es werden jeweils zuerst die Ausnahmen behandelt, die nicht die Satzgrenze markieren; alle betroffenen Stellen müssen einzeln geprüft werden. Ein Satz wird hier als eine typografische Einheit verstanden, für die Kennzeichnung der Satzgrenzen werden Regularitäten im Text verwendet:

Satzanfang:

- am Anfang eines Absatzes
- ein Großbuchstabe (nach einem Satzendezeichen)

Satzende:

- am Ende eines Absatzes
- ein Satzendezeichen (gefolgt von einem Großbuchstaben)

1) Punkt

Punkte, die kein Satzende bezeichnen:

a-1) Punkt nach Ordinalzahl:

Seit dem 21. August verbrachte ich die Tage und Nächte nämlich auf den Straßen, […]

a-2) Punkt nach römischer Zahl:

‚*Zar Nikolaus II.*‘, ‚*III. Klasse*‘

b) Abkürzungen:

‚*k.u.k.-Amtes*‘, ‚*c. k.*‘, ‚*o.k.*‘, ‚*T.G. Masaryk*‘

c) Sonderfälle: Es handelt sich um ein Satzende, aber nach dem Punkt folgt die schließende Klammer, das Satzende-Tag (‚</s>‘) kommt erst nach der Klammer:

Jetzt, wo wir im Bilde sind, kommt es auf das Wort auch nicht an.)

Nachdem alle Ausnahmen behandelt wurden, bezeichnen die übrig gebliebenen Punkte die Satzgrenzen und können als solche automatisch markiert werden.

2) Fragezeichen

Fragezeichen an den Textstellen, wo es sich nicht um eine Satzgrenze handelt:

a) es folgt ein Kleinbuchstabe (direkte Rede):

Was ist das für ein Zirkus? wollte ich wissen.

b) es folgt eine schließende Klammer und Kleinbuchstabe (Einschub):

[…] *dann als Mätresse des Herrn Ingenieurs (denn wie hätte man unser von niemand gesegnetes Verhältnis anders nennen sollen?) und zu guter Letzt als Alžbětkas Stiefmutter.*

c) es folgt eine schließende Klammer und Großbuchstabe:

Představ si, že když čtu třeba (po kolikáté už?) Dostojevského Zločin a trest a […]

d) es folgt Komma und Kleinbuchstabe:

[…] *Alžbětka, která se pak provdala za syna zástupce ředitele vlivné vídeňské banky Creditanstalt?, po válce tedy Sáva nebo Alžbětka emigrovala* […]

e) es folgt eine schließende Klammer, Komma und Kleinbuchstabe (Einschub):

[…] *gerade über ihn hätte sprechen sollen (über wen sonst, wenn nicht über ihn?), tat man aber nicht.*

f) Einschub mit Gedankenstrichen und es folgt ein Kleinbuchstabe:

[…] *seine Frau, will sagen, seine Tochter – sehen Sie, das werde ich ewig durcheinanderbringen, war es Sáva oder Alžbětka, die dann den Sohn des stellvertretenden Direktors der einflußreichen Wiener Creditanstalt geheiratet hat? –, nach dem Krieg* […]

g) es folgt ein Ausrufezeichen:

Was für eine Botschaft?! schreit Angelika.

Fragezeichen in der Funktion eines Satzendezeichens:

a) am Absatzende – wird bei Absatzmarkierungen bearbeitet:

[…] *modernen Zeitalters genannt, geboren?*

[…] *das Transzendente bequem oder gar faul sein?)*

b-1) im Text, es folgt ein Großbuchstabe; es kann auch das Ende einer direkten Rede bezeichnen:

[…] *nicht gesagt, wie das Luftschiff hieß? Also, nun wissen Sie's* […]

b-2) elliptische Sätze:

Krieg? Was für ein Krieg schon wieder? Wovon sprechen Sie denn?

Wie alt war ich jetzt? Sechzehn? Sechzehneinhalb? Also gut, ich zog […]

b-3) es folgt ein Großbuchstabe in Klammern:

Sie fürchten also? (Der SS-Mann konnte sich nicht helfen, aber er fühlte, wie ihm diese dreiste Mißgeburt, die sich fürs […]

3) Ausrufezeichen

Ausrufezeichen an den Textstellen, wo es sich nicht um eine Satzgrenze handelt:

a) es folgt ein Kleinbuchstabe (meistens direkte Rede):

Nicht! schrie ich, nicht aufheben!

b) es folgt die schließende Klammer und Kleinbuchstabe (Einschub), bzw. auch mit Komma:

Nun (und jetzt folgen Sie mir bitte einen Augenblick aufmerksam, es ist wichtig!) sahen wir einen Ausschnitt […]

Mich ärgerte das Merkwürdige, Andersartige an ihm (das harte Ypsilon nach einem weichen Konsonanten!), und ich empfand ihn als eine provokante Häßlichkeit.

c) es folgt ein Geviertstrich, Komma und Kleinbuchstabe:

Es war ein wunderschöner Tag, für einen Ausflug wie geschaffen — dieser unvergeßliche Sommer 1914! —, und alle […]

d) Einschub am Satzende (d. h. Klammer, Punkt gefolgt von einem Großbuchstaben), das Satzendetag (</s>) kommt erst nach dem Punkt:

[…] *ihm jedoch nicht gelang, als hätte es einen Pfropfen in der Kehle (möglicherweise aber auch in der Seele!). Und Flik, der Affe,* […]

e) eine Sonderform:

Wir verbrachten eine Menge Zeit mit diesem Sonja!- und Bruno!-Ding, unsere Namen fielen wie Blütenblätter, […]

Ausrufezeichen in der Funktion eines Satzendezeichens:

a) am Absatzende – wird bei Absatzmarkierungen bearbeitet:

[…] *mich auf ein Treffen mit einem wirklichen Bräutigam vor!*

b) am Absatzende mit einer schließenden Klammer:

[…] *Schreiben hatte er es zwar nicht so, aber wenn es nötig war, zu schreiben, dann schrieb er!)*

c-1) im Text, es folgt ein Großbuchstabe:

Der Schimpanse Flik, das bin ich, dein Bruno! Sonja, du meine Liebe!

c-2) elliptische Sätze:

[…] *nichts übrig. Nichts paßte! Also noch mal! Da begriff ich* […]

c-3) es folgt ein Großbuchstabe nach schließender Klammer (meistens ein Einschub – Kontrolle nötig):

(Hier hättest du dich austoben können, Denis Kotatschka, aufrichtige Grüße in die Hölle!) Und hier siedelte man also […]

aber nach *!)* – eventuell kein Satzende:

[…] *Kanne duftender (stark gewürzter!) Blutsuppe und dem Korb* […]

4) Auslassungszeichen – drei Punkte

Drei Punkte an den Textstellen, wo es sich nicht um eine Satzgrenze handelt:

a) es folgt ein Großbuchstabe, der nicht den Anfang eines neuen Satzes bedeutet:

Wenn Sie nicht auf der Liste der ... Rekruten für das Konzentrationslager stehen, tragen wir Sie eben dort ein.

Mögen Sie recht behalten, Sergeant ... Trotzkij.

b) es folgt ein Kleinbuchstabe:

Ein Mädchen also ... ein Mädchen ... das hat mir niemand gesagt.

Drei Punkte in der Funktion eines Satzendezeichens:

a) am Absatzende:

Unsern guten Kaiser Franz ...

b) es folgt ein Großbuchstabe:

Falls Sie aber jetzt denken, daß damit alles zu Ende war ... Es erwartete mich noch,naja, wie soll man's nennen, […]

Im abschließenden Schritt der Text-Bearbeitung werden die drei Punkte als ein Auslassungszeichen kodiert.

Im bearbeiteten Text kommen keine anderen Satzendezeichen vor. In Abhängigkeit von dem jeweils bearbeiteten Text können Satzendezeichen und andere Zeichen kombiniert werden und die automatische Erkennung als Satzgrenze erschweren: Klammern, Anführungszeichen, Zeilenumbruch oder enthaltene Kodierung der Formatierung.

Quelltext

Die erste Seite der deutschen Fassung des Kratochvil-Textes in dem Zustand, wie sie nach der OCR-Bearbeitung und den folgenden Korrekturen zur weiteren Aufbereitung mit den PHP-Scripts, als eine Textdatei ohne weitere Formatierung (`quelltext_kratochvil_de.txt`), zur Verfügung steht:

```
1
<hi rend="smallCaps">Die Stimme meines Herrn</hi>
XYQGeboren wurde ich, wenn Sie das wirklich hören wol-
len, meine Herren, geboren wurde ich in der Nacht vom
31. Dezember 1899 auf den 1. Januar 1900, und mein Vater
war der Sohn eines orthodoxen russischen Priesters und meine
Mutter eine Deutsche (ihren Eltern — und später ihrem Bru-
der — gehörte ein großes Gut in Landskron am Fuße des
Adlergebirges). Ich wurde zu Hause geboren, damals machte
man das noch so. Meine Hebamme war Magda, eine Ungarin
aus Preßburg. Meine Geburtsstadt Brünn. Die genaue Stelle
dann ein Bett, eine Pritsche im dritten Stock eines großen Miets-
hauses in der Ferdinandstraße (später Masarykstraße, noch spä-
ter Hermann-Göring-Straße und noch später wieder Masaryk-
straße und noch später Straße des Sieges und dann kurz wieder
Masarykstraße und dann wieder lange Zeit Straße des Sieges
und heute wieder Masarykstraße).
XYQJede Geburt ist, falls Sie das nicht wissen sollten, meine Lie-
ben, für alle Beteiligten auch ein Zirkus von Emotionen. Die He-
bamme hielt mich und schrie etwas auf ungarisch und slo-
wakisch. Mutter versuchte, mich anzusehen, und obwohl man sie
daran zu hindern versuchte und mit aller Gewalt ins Bett zurück-
drückte, stützte sie sich auf die Ellenbogen, richtete sich auf wie
eine Kobra und murmelte einen Augenblick hastig etwas auf
tschechisch und gleich darauf wieder auf deutsch. Vater kniete auf
dem Boden, und er, der es mit dem Glauben bisher nicht so genau
genommen hatte, betete jetzt auf russisch, und in diesen großen
orthodoxen Strom, der sich aus ihm herauswälzte, flossen auch
9
```

PHP-Skript

Als Vertreter der verwendeten PHP-Skripte wird hier das Script angeführt, das den deutschen Kratochvil-Text bearbeitet (Datei: `KR_plain_to_xml_kr_de.php`). Die Zeilennummerierung wurde für die Darstellung in diesem Anhang zusätzlich hinzugefügt und gehört nicht zum eigentlichen Dateiinhalt.

```
1     <?php
2     header('content-type: text/html; charset=utf-8');
3     include ("KR_plain_to_xml_funktionen.php");
4     set_time_limit(180);
5     print <<<EndOfHtml
6     <!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
7     <html>
8     <head>
9       <title>PlainToXML - DE</title>
10          <meta    http-equiv="Content-Type"    content="text/html;
charset=utf-8">
11    </head>
12    EndOfHtml;
13
14    //  wenn  es  mit  Linux  gearbeitet  wird,  müssen  die  zu
bearbeitenden
15    // Dateien entsprechende Rechte haben
16
17    // die Quelldatei ('plain text file') darf nicht geändert werden
18    $startdatei = "quelltext_kratochvil_de.txt";
19
20    // die Zieldatei
21    $zieldatei = "zieldatei_KR_de_format_xml.xml";
22
23    // eine eventuelle ältere Version der Zieldatei wird gelöscht
24    if ( file_exists($zieldatei)) {
25        unlink($zieldatei);
26     }
27
28    // Kontrolle: falls es die Quelle nicht gibt, wird das Skript
gestoppt
29    if (! file_exists($startdatei)) {
30         echo "<b>Die Datei</b> \"$startdatei\" <b>existiert <font
size=\"+5\" color=\"red\">NICHT</font></b><br /><br /><br />";
31    } else {
32    // Funktion: Datei wird in ein Array eingelesen und "?", "!" und
"..."
33    // werden entsprechend als Satzgrenze markiert.
34        $fragezeichen = fragezeichen ($startdatei);
35    }
36
37    //  -  das  Ergebnis  der  "Fragezeichen-Funktion"  dient  als
Ausgangsbasis für die weitere Bearbeitung
38    // - neue Variable entsteht, sie dient als Zwischenablage des
39    // Ergebnisses von Kapitelmarkierung
40
41    // aus der Zeile wird der Zeilenumbruch genommen, sonst kommt
"</head>" auf neue Zeile = Fehler
```

```
42                      $muster="#[\r\n]\<hi     rend=\"smallCaps\"\>(?
=.*[\r\n]?.*[\r\n]?.*\<\/hi\>[\r\n])#";
43        $replace="\r\n<head>";
44        $zeile = preg_replace($muster,$replace,$fragezeichen);
45
46        // Zusatz-Korrektur
47        $muster="#<head>(?=Bruno\!)#";
48        $replace='<hi rend="smallCaps">';
49        $zeile = preg_replace($muster,$replace,$zeile);
50
51              // das Tag "</hi>" schließt Attribute sowohl "italic"
als  auch  "smallCaps"  ein  -  alle  Stellen  nach  </head>  müssen
kontrolliert werden (folgt)
52        $muster="#\<\/hi\>(?=[\r\n])#";
53        $replace="</head>";
54        $zeile = preg_replace($muster,$replace,$zeile);
55
56          // Zusatz-Korrektur
57            $muster="#(?<=niščeta,|Drakon|Soničko\!|Evans|Gedärmen\!|
Weltuntergangs\.|ethischen|entgegen.|G\.                M\.|chorchoj\.|
Wahlverwandtschaften,|Wende  herbeizuführen\!|po  cepi  krugom  \.\.\.|
Tiere  nicht  an  sie  verkauft\.|Eisenhower-Doktrin|frißt  Rosen\.|
Bruno\!|Sonja\!|Besatzungsarmeen\.|späči\!|leicht\.)</head>#";
58          $replace="</hi>";
59          $zeile = preg_replace($muster,$replace,$zeile);
60                 // "Träume" kommt einmal in einer Überschrift und
einmal im Satz vor, muss getrennt bearbeitet werden
61                           $muster="#(?<=Träume)</head>[\r\n][\r\n]?(?
=erklang)#";
62               $replace="</hi>\r\n";
63               $zeile = preg_replace($muster,$replace,$zeile);
64
65          $muster="#(?<=es Buch)</head>(?=[\r\n])#";
66          $replace='';
67          $zeile = preg_replace($muster,$replace,$zeile);
68
69          $muster="#(?<=[A-Z][A-Z][A-Z])[\r\n]#";
70          $replace="</head>\r\n";
71          $zeile = preg_replace($muster,$replace,$zeile);
72
73       // ein Array wird gebildet
74         $muster="#(?<=[\r\n]|[\r\n][\r\n])#";      // am Anfang jeder
Zeile steht ein "#" um ein Array nach Zeilen zu bilden
75        $replace="#";
76        $zeile = preg_replace($muster,$replace,$zeile);
77
78              $muster="/#[\r\n]{1,2}(?=#)/";      // am Anfang jeder
Zeile steht ein "#" um ein Array nach Zeilen zu bilden
79        $replace="";
80        $zeile = preg_replace($muster,$replace,$zeile);
81
82        $array = split ( "#", $zeile );
83
84    // bis hierhin wurden die Kapitälchen in <head> umgewandelt, es
folgen die Nummern
85
```

```
86    $kap_z  =  "0";    //  der  Zähler,  nach  dem  die  Kapitelnummer
getestet wird
87
88    foreach ($array as $zeile) {
89      if ( $zeile == $kap_z+1 ) {   // eine Überprüfung mit "is_int"
klappt nicht - wird direkt mit Zaehler verglichen
90
91           $muster="/[\r\n]/";         //  aus  der  Zeile  wird  der
Zeilenumbruch genommen, sonst kommt "</head>" auf neue Zeile = Fehler
92        $replace="";
93        $zeile = preg_replace($muster,$replace,$zeile);
94
95         $muster="#$zeile#";    // nimmt Inhalt von der Zeile -> die
Kpt.-Nr.
96        $replace="<head>$zeile</head>\r\n";
97        $zeile = preg_replace($muster,$replace,$zeile);
98
99        $kap_z++;
100     }
101          $kapitelMarkierung .= $zeile;  // schreibt den Inhalt von
$erg in die neue Variable
102   }
103   ////////////////////////////////////////////////////////
104   // Die Variable, die mit den markierten Kapiteln gebildet wurde,
105   // wird jetzt bearbeitet um die Seitennummerierung zu markieren.
106       // Korrektur - </head><head> wird gelöscht
107       $muster="#</head>[\r\n]<head>#";
108       $replace="\r\n";
109       $zeile = preg_replace($muster,$replace,$kapitelMarkierung);
110         $muster="#</head>[\r\n][\r\n]<head>#";
111         $replace="\r\n";
112         $zeile = preg_replace($muster,$replace,$zeile);
113           $muster="#</head>[\r\n][\r\n][\r\n]<head>#";
114           $replace="\r\n";
115           $zeile = preg_replace($muster,$replace,$zeile);
116
117       // ein Array wird gebildet
118       $muster="#(?<=[\r\n]|[\r\n][\r\n])#";     // am Anfang jeder
Zeile steht ein "#" um ein Array nach Zeilen zu bilden
119       $replace="#";
120       $zeile = preg_replace($muster,$replace,$zeile);
121
122        $muster="/#[\r\n]{1,2}(?=#)/";     // am Anfang jeder Zeile
steht ein "#" um ein Array nach Zeilen zu bilden
123       $replace="";
124       $zeile = preg_replace($muster,$replace,$zeile);
125
126       $zeilens = split ( "#", $zeile );
127
128   // es gibt neue Variable, sie dient als Zwischenablage des
Ergebnisses von
129   // Seitennummern-Markierung, all neuer Text kommt nach hinten
angehängt
130
131   $seite_z = "2";     // Zähler nach dem die Seitennummer getestet
wird
132
```

```
133   //  die  Nummerierung  an  sich  kann  wird  bereits  hier  nicht
verwendet,
134   // weil am Ende muss die Reihenfolge <lb><pb/> in <pb/><lb> aus-
135   // getauscht werden, damit später die Zeilen-Nummerierung in Ab-
136   // hängigkeit von "pb" gezählt werden kann (XSLT).
137
138   foreach ($zeilens as $zeiles) {
139     if ( $zeiles == $seite_z+1 ) {
140
141        $muster="/$zeiles/"; // nimmt Inhalt von der Zeile -> die
Kpt.-Nr.
142          $replace="<pb/>";    //  auf  2  Zeilen  dargestellt,  muss
verbunden wrden
143       $zeiles = preg_replace($muster,$replace,$zeiles);
144
145       $seite_z++;
146     }
147          $seitenMarkierung .= $zeiles;  // schreibt den Inhalt von
$erg in die neue Variable
148   }
149
150   // ************************************************
151   // Funktion zum Austauschen im deutschen Text wird gerufen
152   $erg = austauschen ($seitenMarkierung);
153   // ************************************************
154
155       // ein Array wird gebildet
156       $muster="#(?<=[\r\n]|[\r\n][\r\n])#";     // am Anfang jeder
Zeile steht ein "#" um ein Array nach Zeilen zu bilden
157       $replace="#";
158       $zeile = preg_replace($muster,$replace,$erg);
159
160        $muster="/#[\r\n]{1,2}(?=#)/";      // am Anfang jeder Zeile
steht ein "#" um ein Array nach Zeilen zu bilden
161       $replace="";
162       $zeile = preg_replace($muster,$replace,$zeile);
163
164       $zeilens = split ( "#", $zeile );
165
166   // an den Anfang wird noch Text angehängt
167   $zeileA = array_unshift ( $zeilens, '<TEI><text><body><div>#' );
168
169   // an das Ende wird noch Text angehängt
170   $zeileB = array_push ( $zeilens, '#</body></text></TEI>' );
171
172   $ziel = implode ( '#', $zeilens ); // die Elemente des Array
kommen in eine Variable
173
174   $ziel = preg_replace('/#/','',$ziel);       // die gebliebenen
Verbindungs-Zeichen werdengelöscht
175
176   // ######################################################
177
178   $ziel = zusatz_korrektur_de ($ziel);
179
180   $ziel = synop_de ($ziel);
181
```

```
182   $ziel = sonstiges_de ($ziel);
183
184   // die Kapiteln kommen in ein "div" für Kapitel
185
186   $muster="#(?<=[\r\n])(?=<pb/><lb/><head>[0-9])#";
187   $replace='</chapter><chapter type="chapter">';
188   $ziel = preg_replace($muster,$replace,$ziel);
189
190   // Ergänzung: Anfang des Kapitel-DIV auf eigene Zeile
191   $muster="#(?<=\</chapter\>)(?=\<chapter type=\"chapter\"\>)#";
192   $replace="\r\n";
193   $ziel = preg_replace($muster,$replace,$ziel);
194
195   // Korrektur der jeweils ersten Kapitelmarkierung (in einem
Buch)
196   $muster="#</head>[\r\n][\r\n]?<pb/>[\r\n][\r\n]?</chapter>#";
197   $replace="</head>\r\n<pb/>";
198   $ziel = preg_replace($muster,$replace,$ziel);
199
200         // Korrektur: Erg. der abschließenden Kapitelmarkierung am
Romanende
201         $muster="#(?=</div></body>)#";
202         $replace="</chapter>\r\n";
203         $ziel = preg_replace($muster,$replace,$ziel);
204
205         // Korrektur - Beenden des jeweils letzten Kapitels in
einem Buch
206         $muster="#</div>[\r\n][\r\n]?(?=<div>)#";
207         $replace="</chapter></div>\r\n";
208         $ziel = preg_replace($muster,$replace,$ziel);
209
210   $muster="#<div>#";
211   $replace='<div type="book">';
212   $ziel = preg_replace($muster,$replace,$ziel);
213
214   /////// Korrektur der falsch platzierten Seitenumbrüche
215
216   $muster="#<pb/><lb/><head>(?=2[\r\n][\r\n]?)#";
217   $replace='<lb/><head>';
218   $ziel = preg_replace($muster,$replace,$ziel);
219
220   $muster="#<pb/><lb/><head>(?=5[\r\n][\r\n]?)#";
221   $replace='<lb/><head>';
222   $ziel = preg_replace($muster,$replace,$ziel);
223
224   $muster="#<pb/><lb/><head>(?=6[\r\n][\r\n]?)#";
225   $replace='<lb/><head>';
226   $ziel = preg_replace($muster,$replace,$ziel);
227
228   $muster="#<pb/><lb/><head>(?=8[\r\n][\r\n]?)#";
229   $replace='<lb/><head>';
230   $ziel = preg_replace($muster,$replace,$ziel);
231
232   $muster="#<pb/><lb/><head>(?=9[\r\n][\r\n]?)#";
233   $replace='<lb/><head>';
234   $ziel = preg_replace($muster,$replace,$ziel);
235
```

```
236
237   $muster="#<pb/><lb/><head>(?=10|11|12|13|15|16|20|21|22|23|26|
27|29|32|33|35|38|39|40|41|43|44|45|48|49|50|51|52|53|56|59|60|64|65|
66)#";
238   $replace='<lb/><head>';
239   $ziel = preg_replace($muster,$replace,$ziel);
240
241   $muster="#(?<=Jahrtausendwasser)(?=</hi>)#";
242   $replace='.';
243   $ziel = preg_replace($muster,$replace,$ziel);
244
245   $ziel = preg_replace("/\.\.\./","…",$ziel);   // die drei Punkte
werden in Auslassungszeichen umgewandelt
246   // ######################################################
247   // die römischen Zahlen sind im Buch in Kapitälchenschrift -
Tagging wird ergänzt
248   // i. - ic - ii. - iii. - iv. - xvi.
249   $muster="#(?<= )I\.#";
250   $replace='<hi rend="smallCaps">i</hi>.';
251   $ziel = preg_replace($muster,$replace,$ziel);
252
253   $muster="#(?<= )II\.#";
254   $replace='<hi rend="smallCaps">ii</hi>.';
255   $ziel = preg_replace($muster,$replace,$ziel);
256
257   $muster="#III\.#";
258   $replace='<hi rend="smallCaps">iii</hi>.';
259   $ziel = preg_replace($muster,$replace,$ziel);
260
261   $muster="#(?<= )IV\.#";
262   $replace='<hi rend="smallCaps">iv</hi>.';
263   $ziel = preg_replace($muster,$replace,$ziel);
264
265   $muster="#(?<= )XVI\.#";
266   $replace='<hi rend="smallCaps">i</hi>.';
267   $ziel = preg_replace($muster,$replace,$ziel);
268
269   $muster="#(?<= )Ic\)#";
270   $replace='<hi rend="smallCaps">i</hi>c';
271   $ziel = preg_replace($muster,$replace,$ziel);
272   // ######################################################
273   // Erstellen der Zieldatei
274   $fp = fopen ( $zieldatei, 'a+' );    // öffnet neue Datei zum
Schreiben
275   fwrite ( $fp, $ziel ); // schreibt den Inhalt von $erg in die
Zieldatei
276   fclose ( $fp );
277
278   echo "
279   <table border=\"1\" cellpadding=\"20\"><tr><td>
280   Zieldatei ($zieldatei) erstellt,<br/>\r\n
281   Textkodierung: UFT-8<br/>\r\n
282   </td></tr></table>";
283   ?></body></html>
```

PHP-Funktionen

Die PHP-Skripte werden durch die Datei, in der die PHP-Funktionen gespeichert sind, begleitet (`KR_plain_to_xml_funktionen.php`). Da diese Datei Funktionen sowohl für den deutschen als auch für den tschechischen Kratochvil-Text enthalten, werden hier aus platzsparenden Gründen die Teile, die nur den tschechischen Text betreffen, ausgelassen. Von der Darstellung werden auch Schritte ausgelassen, die die Synoptisierung der Sätze betreffen – das Prinzip wird einigen Beispielen gezeigt, die anderen werden ausgelassen. Die Zeilennummerierung wurde für die Darstellung in diesem Anhang zusätzlich hinzugefügt und gehört nicht zum eigentlichen Dateiinhalt.

```
1     <?
2     header('content-type: text/html; charset=utf-8');
3     function fragezeichen ($startdatei) {
4
5     $fp = fopen ( $startdatei, 'r' );
6     while ( ! feof ( $fp ) ) {
7           $zeilen[] = fgets ( $fp );
8           }
9           fclose ( $fp );
10
11    $array_num ="0";
12    foreach ($zeilen as $zeile) {
13    $array_num++;
14
15    // die Sequenz "?!" wird erst bei Ausrufezeichen behandelt
16      $muster="/\?\!/";
17      $replace="_?_!";
18      $erg = preg_replace($muster,$replace,$zeile);
19    // es folgt ein Großbuchstabe (neue Sätze auch eliptische Sätze)
20      $muster="/\? (?=[A-ZÄÖÜÁČĎÉĚÍŇÓŘŠŤÚŮÝŽ])/";
21      $replace="?</s></seg> <seg><s>";
22      $erg = preg_replace($muster,$replace,$erg);
23    // es folgt ein Großbuchstabe in öff. Klammern
24      $muster="/\? (?=\([A-ZÄÖÜÁČĎÉĚÍŇÓŘŠŤÚŮÝŽ])/";
25      $replace="?</s></seg> <seg><s>";
26      $erg = preg_replace($muster,$replace,$erg);
27    // KEIN SATZENDE: es folgt ein Kleinbuchstabe
28      $muster="/\?(?= [a-zäöüáčďéěíňóřšťúůýž])/";
29      $replace="_?_";
30      $erg = preg_replace($muster,$replace,$erg);
31        $erg = trim ( $erg );   // entfernt Zeilenumbruch
32        $paar = "$erg" . "@" . "$zeilen[$array_num]";
33    // ein "?" am Zeilenende, kommt ein Gr.-B. am Anfang nächster
      Zeile?
34      $muster="/\?(?=@[A-ZÄÖÜÁČĎÉĚÍŇÓŘŠŤÚŮÝŽ])/";
35      $replace="?</s></seg> <seg><s>";
36      $erg = preg_replace($muster,$replace,$paar);
37        $teil = explode("@", $erg);
38        $zeile_neu = $teil[0] . "\n";
39    // es folgt ein Kleinbuchstabe (auf einer neuen Zeile)
40      $erg = trim ( $zeile_neu );   // entfernt Zeilenumbruch
```

```
41      $paar = "$erg" . "@" . "$zeilen[$array_num]";
42    // ein "?" am Zeilenende, kommt ein Kl.-B. am Anfang nächster
      Zeile?
43      $muster="/\?(?=@[a-zäöüáčďéěíňóřšťúůýž])/";
44      $replace="_?_";
45      $erg = preg_replace($muster,$replace,$paar);
46        $teil = explode("@", $erg);
47        $zeile_neu = $teil[0] . "\n";
48    // es folgt Komma und Kleinbuchstabe
49      $muster="/\?(?=, [a-zäöüáčďéěíňóřšťúůýž])/";
50      $replace="_?_";
51      $erg = preg_replace($muster,$replace,$zeile_neu);
52    //  KEIN  ZEILENENDE:    es  folgt  eine  schl.  Klammer  und
      Kleinbuchstabe
53      $muster="/\?(?=\) [a-zäöüáčďéěíňóřšťúůýž])/";
54      $replace="_?_";
55      $erg = preg_replace($muster,$replace,$erg);
56    // KEIN ZEILENENDE:  es folgt eine schl. Klammer, Komma und Kl.-
      B.
57      $muster="/\?(?=\), [a-zäöüáčďéěíňóřšťúůýž])/";
58      $replace="_?_";
59      $erg = preg_replace($muster,$replace,$erg);
60    // KEIN ZEILENENDE:  es folgt ein  Gedankenstrich (—), Komma und
      Kl.-B.
61      $muster="/\?(?= -,)/";
62      $replace="_?_";
63      $erg = preg_replace($muster,$replace,$erg);
64    // KEIN ZEILENENDE:  es folgt ein  Gedankenstrich, Komma und
      Kl.-B.
65      $muster="/\?(?=\.)/";
66      $replace="_?_";
67      $erg = preg_replace($muster,$replace,$erg);
68    ////////////////////////////// AUSRUFEZEICHEN
69    // es folgt ein Großbuchstabe (neue Sätze auch eliptische Sätze)
70      $muster="/\!(?= [A-ZÄÖÜÁČĎÉĚÍŇÓŘŠŤÚŮÝŽ])/";
71      $replace="!</s></seg> <seg><s>";
72      $erg = preg_replace($muster,$replace,$erg);
73    // es folgt ein Großbuchstabe (auf einer neuen Zeile)
74    // Überprüfung von "!" am Zeilenende, ob ein Gr.-B. danach
      vorkommt?
75      $erg = trim ( $erg );   // entfernt Zeilenumbruch
76      $paar = "$erg" . "@" . "$zeilen[$array_num]";
77    // ein "?" am Zeilenende, kommt ein Gr.-B. am Anfang nächster
      Zeile?
78      $muster="/\!(?=@[A-ZÄÖÜÁČĎÉĚÍŇÓŘŠŤÚŮÝŽ])/";
79      $replace="!</s></seg> <seg><s>";
80      $erg = preg_replace($muster,$replace,$paar);
81        $teil = explode("@", $erg);
82        $erg = $teil[0] . "\n";
83        //echo $array_num . $zeile_neu . "\n";
84    // es folgt ein Großbuchstabe nach schl. Klammer
85      $muster="/\!\)(?= [A-ZÄÖÜÁČĎÉĚÍŇÓŘŠŤÚŮÝŽ])/";
86      $replace="_!_)</s></seg> <seg><s>";
87      $erg = preg_replace($muster,$replace,$erg);
88    // KEIN SATZENDE: es folgt ein Gr.-B. nach schl. Klammer mit
      Punkt
89      $muster="/\!(?=\)\.?)/";
```

```
90      $replace="_!_";
91      $erg = preg_replace($muster,$replace,$erg);
92    // Großbuchstabe + Markierung
93      $muster="/\!\<\/\w{1,2}\> (?=[A-ZÄÖÜÁČĎÉĚÍŇÓŘŠŤÚŮÝŽ])/";
94      $replace="!</hi></s></seg> <seg><s>";
95      $erg = preg_replace($muster,$replace,$erg);
96    // Großbuchstabe + Markierung
97    // xxx! <hi rend="italic">Xxx
98                $muster="/\!(?=      \<hi      rend=\"italic\"\>[A-
      ZÄÖÜÁČĎÉĚÍŇÓŘŠŤÚŮÝŽ])/";
99      $replace="!</s></seg> <seg><s>";
100     $erg = preg_replace($muster,$replace,$erg);
101   // KEIN SATZENDE: Kleinbuchstabe + Markierung   (Komma kann
      vorkommen)
102     $muster="/\!(?=\[\/\w{1,}\],? [a-zäöüáčďéěíňóřšťúůýž])/";
103     $replace="_!_";
104     $erg = preg_replace($muster,$replace,$erg);
105   // KEIN SATZENDE: es folgt ein Kleinbuchstabe (Komma kann
      vorkommen)
106     $muster="/\!(?=,? [a-zäöüáčďéěíňóřšťúůýž])/";
107     $replace="_!_";
108     $erg = preg_replace($muster,$replace,$erg);
109   // es folgt ein Kleinbuchstabe (auf einer neuen Zeile)
110     $erg = trim ( $erg );   // entfernt Zeilenumbruch
111     $paar = "$erg" . "@" . "$zeilen[$array_num]";
112   // ein "?" am Zeilenende, kommt ein Kl.-B. am Anfang nächster
      Zeile?
113     $muster="/\!(?=@[a-zäöüáčďéěíňóřšťúůýž])/";
114     $replace="_!_";
115     $erg = preg_replace($muster,$replace,$paar);
116       $teil = explode("@", $erg);
117       $erg = $teil[0] . "\n";
118   // es folgt schl. Klammer, Komma und auf einer neuen Zeile ein
      Kl.-B.
119     $erg = trim ( $erg );   // entfernt Zeilenumbruch
120     $paar = "$erg" . "@" . "$zeilen[$array_num]";
121   // ein "?" am Zeilenende, kommt ein Kl.-B. am Anfang nächster
      Zeile?
122     $muster="/\!(?=\),@[a-zäöüáčďéěíňóřšťúůýž])/";
123     $replace="_!_";
124     $erg = preg_replace($muster,$replace,$paar);
125       $teil = explode("@", $erg);
126       $erg = $teil[0] . "\n";
127   //  KEIN ZEILENENDE: es folgt eine schl. Klammer und
      Kleinbuchstabe
128   //  Variante:    es folgt eine schl. Klammer, Komma und
      Kleinbuchstabe
129     $muster="/\!(?=\),? [a-zäöüáčďéěíňóřšťúůýž])/";
130     $replace="_!_";
131     $erg = preg_replace($muster,$replace,$erg);
132   // KEIN ZEILENENDE:  es folgt ein  Gedankenstrich, Komma und
      Kl.-B.
133     $muster="/\!(?= –, [a-z])/";
134     $replace="_!_";
135     $erg = preg_replace($muster,$replace,$erg);
136   // Korrektur
137     $muster="/(?<=<s>) /";
```

```
138     $replace="";
139     $erg = preg_replace($muster,$replace,$erg);
140   ////////////////////// drei Punkte
141   // es folgt ein Großbuchstabe (auf einer neuen Zeile)
142   // Überprüfung von drei Punkten am Zeilenende, ob Gr.-B. danach
      steht?
143   $erg = trim ( $erg );   // entfernt Zeilenumbruch
144   $paar = "$erg" . "@" . "$zeilen[$array_num]";
145   // ein "?" am Zeilenende, kommt ein Gr.-B. am Anfang nächster
      Zeile?
146     $muster="/\.\.\.(?= ?@ ?[A-ZÄÖÜÁČĎÉĚÍŇÓŘŠŤÚŮÝŽ])/";
147     $replace="...</s></seg> <seg><s>";
148     $erg = preg_replace($muster,$replace,$paar);
149       $teil = explode("@", $erg);
150       $erg = $teil[0] . "\n";
151   $ergebnisDerFunktion .= $erg;
152   }
153   return $ergebnisDerFunktion;
154   }
155
156   function austauschen ($seitenMarkierung) {
157       $muster="#(?<=<\/head>)[\r\n][\r\n]?XYQ#";
158       $replace="\r\n<p><seg><s>";
159       $erg = preg_replace($muster,$replace,$seitenMarkierung);
160         $muster="#[\r\n][\r\n]?XYQ#";
161         $replace="</s></seg></p>\r\n<p><seg><s>";
162         $erg = preg_replace($muster,$replace,$erg);
163           $muster="#[\r\n][\r\n]?<pb/>XYQ#";
164           $replace="</s></seg></p>\r\n<pb/><p><seg><s>";
165           $erg = preg_replace($muster,$replace,$erg);
166   //  Ersetzen  von  Punkten,  die  nicht  als  Satzendemarkierung
      funktionieren
167
168   //  die  <head>-  und  </head>  sind  jetzt  durch  Zeilenumbruch
      getrennt - die
169   //  neuentstandene  Zeile  ist  nicht  wünschenswert  -  muss
      korrigiert werden
170     $muster="/[\r\n](?=<\/head>)/";         // Absatzgrenze;
171     $replace="";
172     $erg = preg_replace($muster,$replace,$erg);
173   // drei Punkte wurden in ein Zeichen umgewandelt - nicht mehr
      gebraucht
174     $erg = preg_replace("/\.\.\./","###",$erg);
175   // drei Punkte werden in ### umgewandelt, damit sie die Muster
      nicht
176   // verwirren + am Ende zurück zu drei Punkten
177      $erg = preg_replace("/\. (?=Januar|Led|led|Februar|Únor|únor|
März|Břez|břez|April|Dub|dub|Mai|Květ|květ|Juni|Juli|Červ|červ|August|
Srp|srp|Septem|Zář|zář|Oktober|Říj|říj|November|Listo|listo|Dezember|
Prosi|prosi)/","# ",$erg);
178     $erg = preg_replace("/\. (?=Jahr)/","# ",$erg);
179   // Zahl+Jahr/hundert/s, wie z.B. : "im 6. Jahrhundert"
180     $erg = preg_replace("/\. (?=[a-z])/","# ",$erg);
181   // Punkt-Blank-Kleinbuchstabe
182   // Abkürzung "k.u.k."  und "o.k."
183     $erg = preg_replace("/k\.u\.k\./","k#u#k#",$erg);
184     $erg = preg_replace("/o\.k\./","o#k#",$erg);
```

```
185
186   // Punkt nach römisch Zahl (bzw. auch mit Komma gefolgt),
      nachdem ein
187   //Kleinbuchstabe kommt wie "Nikolaus II. eine"
188   $muster = "/(?<=I|V|X|L|C|D|M)\.(?= [a-záčďéěíňóřšťúůýäöü])/";
189   $replace = "#";
190   $erg = preg_replace($muster,$replace,$erg);
191   // falls nötig, im gleichen Schritt auch die Gr.-B. nach der
192   // röm. Zahl zu behandeln, müssen zuerst die Kl.-B. bearbeitet
193   //  werden, sonst klappt es mit den akzentuierten Zeichen nicht
194   $muster = "/(?<=I|V|X|L|C|D|M)\.,(?= [a-záčďéěíňóřšťúůýäöü])/";
195   $replace = "#,";
196   $erg = preg_replace($muster,$replace,$erg);
197     $erg = preg_replace("/\.\)/","#)",$erg);
198   // Punkt-vor-Klammer.zu - Ende des Satzes erst nach Klammer?
199   //    besondere    Satzgrenze:    Punkt+Klammer_zu    nun    Fis-
      Zeichen+Klammer_zu
200   //-> wo wird Satzendezeichen gesetzt?
201     $erg = preg_replace("/(?<=[A-Z])\./","#",$erg);
202   // Punkt-nach-Großbuchstabe - wie: Präsidenten T.G. Masaryk
203   // Korrektur der Reihenfolge
204     $muster="/<pb\/>[\r\n][\r\n]?<\/s><\/seg><\/p>/";
205     $replace="</s></seg></p>\r\n<pb/>";
206     $erg = preg_replace($muster,$replace,$erg);
207       $muster="/<pb\/>[\r\n][\r\n]?/";
208       $replace="<pb/>";
209       $erg = preg_replace($muster,$replace,$erg);
210         $muster="/<pb\/>[\r\n][\r\n]?/";
211         $replace="<pb/>";
212         $erg = preg_replace($muster,$replace,$erg);
213       // Ergänzung zu der Reihenfolge
214         $muster="/[\r\n][\r\n]?<\/s><\/seg><\/p>/";
215         $replace="</s></seg></p>";
216         $erg = preg_replace($muster,$replace,$erg);
217         // Satzendepunkte mit </s> kombiniert (wie bei ...) -
      ausgeblendet
218         $muster="/\.<\/s><\/seg><\/p>/";
219         $replace="#</s></seg></p>";
220         $erg = preg_replace($muster,$replace,$erg);
221
222   // die übrigen Satzgrenzen ////////////////  Punkt
223     $muster="/\. (?=\S)/";
224     $replace="#</s></seg> <seg><s>";
225     $erg = preg_replace($muster,$replace,$erg);
226   // Punkt am Zeilenende (wie: "die Liebe bis in den Tod.");
227   // gleichzeitig ausgeblendet, erleichtert weitere Verarbeitung
228     $muster="/\.[\r\n][\r\n]?/";
229     $replace="#</s></seg>\r\n<seg><s>";
230     $erg = preg_replace($muster,$replace,$erg);
231   // Sonderfall - Satzanfang : erst nach der Klammer? wie: "fuhr.
      (Jawohl"
232   // im folgenden Schritt werden alle Punkte wiederhergestellt
233   //- bis jetzt als Fis-Zeichen
234     $erg = preg_replace("/#/",".",$erg);
235   ////////////////////////////////////////////// Zeilenumbruch
236     $muster="/[\r\n]/";
237     $replace="\r\n<lb/>";
```

```
238     $erg = preg_replace($muster,$replace,$erg);
239   // Korrektur
240     $muster="/<lb\/>[\r\n][\r\n]<lb\/>/";
241     $replace="<lb/>";
242     $erg = preg_replace($muster,$replace,$erg);
243   // Reihenfolge-Korrektur
244     $muster="/<lb\/><pb\/>/";
245     $replace="<pb/><lb/>";
246     $erg = preg_replace($muster,$replace,$erg);
247   //  ergänzt  ein  "Linebreak"  zu  dem  ersten  "Pagebreak"  im
      Dokument; v.a.
248   // jeweils die erste Zeile eines Abschnittes
249     $muster="/<pb\/>(?=\w)/";
250     $replace="<pb/><lb/>";
251     $erg = preg_replace($muster,$replace,$erg);
252   return $erg;
253   }
254
255   function zusatz_korrektur_de ($ziel) {
256   //  die  Zeilenumbruchkodierungen  \r  und  \r\n  vereinheitlicht
      in \r\n
257     $muster="#\r(?!\n)#";
258     $replace="\r\n";
259     $datei_inhalt = preg_replace($muster,$replace,$ziel);
260   // Markierung der Überschriften die in Gr.-B. geschrieben sind
261 // der zweite Teil der Überschrift ist in Gr.-B. geschrieben
262           // der zweite Teil der Überschrift ist in Gr.-B. geschr.
263                           $muster="#(?<=[A-ZÁČĎÉĚÍŇÓŘŠŤÚŮÝŽÄÖÜ][A-
ZÁČĎÉĚÍŇÓŘŠŤÚŮÝŽÄÖÜ][A-ZÁČĎÉĚÍŇÓŘŠŤÚŮÝŽÄÖÜ])[\r\n][\r\n]?[\r]#";
264           $replace="</head>";
265           $erg = preg_replace($muster,$replace,$datei_inhalt);
266   // die überflüssigen Zeilenumbrüche werden gelöscht
267         $muster  =  "#<pb/><pb/><pb/><pb/><pb/><pb/><pb/><pb/><pb/>|
<pb/><pb/><pb/><pb/><pb/><pb/><pb/><pb/>|
<pb/><pb/><pb/><pb/><pb/><pb/><pb/>|<pb/><pb/><pb/><pb/><pb/><pb/>|
<pb/><pb/><pb/><pb/><pb/>|<pb/><pb/><pb/><pb/>|<pb/><pb/><pb/>#";
268       $replace = "<pb/>";
269       $erg = preg_replace($muster,$replace,$erg);
270       // am Ende der Datei
271         $muster = "#<pb/><lb/>(?=</body>)#";
272         $replace = "";
273         $erg = preg_replace($muster,$replace,$erg);
274   //Behandlung der Umgebung von <head>
275       $muster = "#<pb/><lb/><pb/><lb/>#";
276       $replace = "<pb/><lb/>";
277       $erg = preg_replace($muster,$replace,$erg);
278         $muster = "#<pb/><lb/><pb/>#";
279         $replace = "<pb/><lb/>";
280         $erg = preg_replace($muster,$replace,$erg);
281   // Zeichenfolge "XYQ", die den Absatzanfang markiert darf sich
282       // nicht innerhalb eines <head>-Elements befinden
283         $muster = "#(?<=<head>)XYQ#";
284         $replace = "";
285         $erg = preg_replace($muster,$replace,$erg);
286   // <body>
287                           $muster   =   "#(?<=<body>|<lb/>)[\r\n]?
[\r\n]?.?.?.?.?.?.?.?.?2#";
```

```
288           $replace = "";
289           $erg = preg_replace($muster,$replace,$erg);
290       //  Löschen  der  überflüssigen  Elemente  +  Korrektur  der
      Reihenfolge
291       $muster = "#
292                </s></seg>[\r\n][\r\n]?<lb/><seg><s><pb/><pb/><lb/>(?
=<head>)|</s></seg>[\r\n][\r\n]?<lb/><seg><s><pb/><lb/>(?=<head>)|
</s></seg>[\r\n][\r\n]?<lb/><seg><s><pb/>(?=<head>)|</s></seg>[\r\n]
[\r\n]?<lb/><seg><s>(?=<head>)#";
293       $replace = "</s></seg></p>\r\n<pb/><lb/>";
294       $erg = preg_replace($muster,$replace,$erg);
295     // zusätzlich wird das fehlende Abs.-Ende für die vorh. Zeile
      ersetzt
296           // Teil a)  :  "?" "!" ")" vor "<lb/><head>""
297           $muster = "#(?<=\?|\!|\))[\r\n][\r\n]?<lb/>(?=<head>)#";
298           $replace = "</s></seg></p>\r\n<pb/><lb/>";
299           $erg = preg_replace($muster,$replace,$erg);
300           // Teil b)  :  "?" "!" ")" vor "<pb/><lb/><head>""
301                $muster  =  "#(?<=\?|\!|\))[\r\n][\r\n]?<pb/><lb/>(?
=<head>)#";
302           $replace = "</s></seg></p>\r\n<pb/><lb/>";
303           $erg = preg_replace($muster,$replace,$erg);
304      // Ergänzen von Leerseiten - die sind wegen Seitennummer.
      einzusetzen
305       // nach diesen Regeln:
306           // nach jedem "Buch" kommt eine Leerseite.
307           $muster = "#</head>(?=[\r\n]?[\r\n]?<pb/><lb/><head>)#";
308           $replace = "</head>\r\n<pb/>";
309           $erg = preg_replace($muster,$replace,$erg);
310     // Jedes "Buch" wird in ein "<div>" positioniert
311                $muster  =  "#<pb/><lb/><head>(?=Zweites  Buch|Drittes
Buch|Viertes Buch|Fünftes Buch)#";
312           $replace = "</div>\r\n<div>\r\n<pb/><lb/><head>";
313           $erg = preg_replace($muster,$replace,$erg);
314         // Korrektur: vor dem ersten <div> darf kein </div> stehen
315             $muster = "#<pb/><lb/><head>(?=Erstes Buch)#";
316             $replace = "<div>\r\n<pb/><lb/><head>";
317             $erg = preg_replace($muster,$replace,$erg);
318           // Ergänzung: das letzte <div> wird geschlossen
319             $muster = "#</body>#";
320             $replace = "</div></body>";
321             $erg = preg_replace($muster,$replace,$erg);
322             // Ergänzung: das letzte <div> wird geschlossen
323                                $muster  =  "#.</s></seg>[\r\n][\r\n]?
<lb/><seg><s><pb/>(?=</div></body>)#";
324                $replace = "</s></seg></p>\r\n";
325                $erg = preg_replace($muster,$replace,$erg);
326       // "<head>" befindet sich nicht in einem Absatz, dies
327       // "</head></s></seg></p>" muss korrigiert werden
328           $muster = "#(?<=</head>)</s></seg></p>#";
329           $replace = "";
330           $erg = preg_replace($muster,$replace,$erg);
331   // andere Sonderfälle
332       // leerer Absatz " <p><seg><s></s></seg></p>"
333                          $muster  =  "#  ?<p><seg><s></s></seg></p>|
<lb/><p><seg><s><pb/>#";
334           $replace = "";
```

```
335           $erg = preg_replace($muster,$replace,$erg);
336       // leeres Segment " <seg><s> ?</s></seg>"
337           $muster = "# ?<seg><s> ?</s></seg>#";
338           $replace = "";
339           $erg = preg_replace($muster,$replace,$erg);
340       // die übriggebliebenen Absatzanfänge (XYQ)
341           // Reihenfolge
342                                $muster   =   "#</s></seg>[\r\n][\r\n]?
<lb/><seg><s><pb/><lb/>XYQ#";
343           $replace = "</s></seg></p>\r\n<pb/><lb/><p><seg><s>";
344           $erg = preg_replace($muster,$replace,$erg);
345           // die restlichen Absatzanfänge am Zeilenumbruch
346          // zwei davon sind keine Satzenden, korrigieren!
347               $muster = "#XYQ(?=gedrückt|Zusammentreffen)#";
348               $replace = "";
349               $erg = preg_replace($muster,$replace,$erg);
350            // die anderen können mit Absatzmarken versehen werden
351               // Muster: "?" "!" ")" vor <pb/><lb/>XYQ
352                                $muster  =  "#(?<=\?|\!\))[\r\n][\r\n]?
<pb/><lb/>XYQ#";
353                                                          $replace    =
"</s></seg></p>\r\n<pb/><lb/><p><seg><s>";
354                 $erg = preg_replace($muster,$replace,$erg);
355               $muster = "#[\r\n][\r\n]?<pb/><lb/>XYQ(?=Währen|Ich|
Weiter)#";
356             $replace = "</s></seg></p>\r\n<pb/><lb/><p><seg><s>";
357             $erg = preg_replace($muster,$replace,$erg);
358                        // Muster: <pb/><lb/>XYQ
359                        $muster = "#(?<=[\r\n])<lb/>XYQ#";
360                        $replace = "<lb/><p><seg><s>";
361                        $erg = preg_replace($muster,$replace,$erg);
362       // Datei überflüssiger leerer Absatz
363         $muster = "#<lb/><p><seg><s><pb/></s></seg></p>\r\n#";
364         $replace = "";
365         $erg = preg_replace($muster,$replace,$erg);
366       // Brief : der Brief wurde noch nicht beendet:
367         // nach: <lb/>Thomas G. M.</hi>
368         $muster = "#[\r\n][\r\n]?(?=<pb/><lb/><head>34)#";
369         $replace = "</s></seg></p>\r\n";
370         $erg = preg_replace($muster,$replace,$erg);
371          // überflüssige Kursiv-Ende- und Kursiv-Anfang-Markierung
      im Brief
372         $muster = "#(?<=einzigen ethischen)</hi>#";
373         $replace = "";
374         $erg = preg_replace($muster,$replace,$erg);
375           $muster = "#<hi rend=\"italic\">(?=Überzeugung)#";
376           $replace = "";
377           $erg = preg_replace($muster,$replace,$erg);
378       // & -> &amp;
379         $muster = "#&#";
380         $replace = "&amp;";
381         $erg = preg_replace($muster,$replace,$erg);
382       // mehrere Absatzende-Markierungen hintereinander
383            $muster  =  "#</s></seg></p></s></seg></p></s></seg></p>|
</s></seg></p></s></seg></p>#";
384         $replace = "</s></seg></p>";
385         $erg = preg_replace($muster,$replace,$erg);
```

```
386       // es fehlt Seitenumbruch
387             $muster = "#(?<=sei Ekstase ...</s></seg>)</p>[\r\n]
[\r\n]?<lb/>#";
388         $replace = "\r\n<pb/><lb/><seg><s>";
389         $erg = preg_replace($muster,$replace,$erg);
390   // Reihenfolge bei Satzende wo "<p><s><seg>" vor "<lb/>" steht
391         $muster = "# <p><seg><s>[\r\n][\r\n]?<lb/>#";
392         $replace = "\r\n<lb/><p><seg><s>";
393         $erg = preg_replace($muster,$replace,$erg);
394   // Korrektur: Teile der direkten Rede werden verbunden
395   $erg = preg_replace('#(?<=macht schon\!)</s></seg> <seg><s>(?
=Worauf)#',' ',$erg);
396   $erg = preg_replace('#(?<=wartet ihr\?) ?</s></seg> ?
<seg><s>#','',$erg);
397   $erg = preg_replace('#<seg><s>(?=Und was machst du)#',' ',$erg);
398   $erg = preg_replace('#(?<=noch mal\!)</s></seg> <seg><s>(?
=Zieh)#',' ',$erg);
399   $erg = preg_replace('#(?<=ihr raus\!)</s></seg> <seg><s>(?=Jetzt
mußt)#',' ',$erg);
400   $erg = preg_replace('#(?<=drinlassen\!)</s></seg> <seg><s>(?=Du
willst der)#',' ',$erg);
401   $erg = preg_replace('#</s></seg> <seg><s>(?=Also, stoß zu,)#','
',$erg);
402   $erg = preg_replace('#(?<=siehst du\!)</s></seg> <seg><s>(?
=Ahhh)#',' ',$erg);
403   // Korrektur
404         $erg = preg_replace('#(?<=Seit dem 21\.)</s></seg>
<seg><s>#',' ',$erg);
405   // Korrektur
406     $muster = "#<pb/><div>#";
407     $replace = "<div>";
408     $erg = preg_replace($muster,$replace,$erg);
409   ////////////////////////////////////
410   $erg = preg_replace('#pozor\!</s></seg> <seg><s>Ostorožno\!
</s></seg> <seg><s>Ostorožno\!#','pozor! Ostorožno! Ostorožno!',$erg);
411      return $erg;
412   }
413
414   function zusatz_korrektur_cz ($ziel) {
[...]
551      return $erg;
552   }
553
554   function synop_de ($ziel) {
555   $erg           =           preg_replace('#(?<=leidenschaftlicher
Naturkundler\.</s>)</seg>     <seg><s>(?=Er     überlegte)#','     <s
attr="2_1">',$ziel);
556   $erg = preg_replace('#(?<=</s>)</seg> <seg><s>(?=Sehen Sie, wie
die Tiere da springen)#',' <s attr="2_1">',$erg);
557   $erg = preg_replace('#(?<=zurückbringen würde\.</s>)</seg>
<seg><s>(?=Diese Last fiel)#',' <s attr="2_1">',$erg);
[...]
775     $erg = preg_replace('# attr="\d_\d"#','',$erg);
776   return $erg;
777   }
778
779   function synop_cz ($ziel) {
```

```
[...]
845     $erg = preg_replace('# attr="\d_\d"#','',$erg);
846   return $erg;
847   }
848
849   function sonstiges_de ($ziel) {
850   // S. 61
851   $erg = preg_replace('#Unsern guten Kaiser Franz\!<\/s><\/seg>
<seg><s>#','<l><seg>Unsern guten Kaiser Franz!</seg></l>',$ziel);
[...]
914   return $erg;
915   }
916
917   function sonstiges_cz ($erg_sonstiges) {
[...]
977   return $erg;
978   }
979   ?>
```

Der Text in grober XML-Struktur

Die erste Seite der deutschen Fassung des Kratochvil-Textes in der Form, wie die Daten mit den XML-Tags versehen wurden, bevor die endgültige Taganpassung und -Attribuierung mit der XSLT-Schablone durchgeführt wird, es handelt sich um die Datei: `zieldatei_KR_de_format_xml.xml`

```
<chapter type="chapter"><pb/><lb/><head>1
<lb/>Die Stimme meines Herrn</head>
<lb/><p><seg><s>Geboren wurde ich, wenn Sie das wirklich hören wol-
<lb/>len, meine Herren, geboren wurde ich in der Nacht vom
<lb/>31. Dezember 1899 auf den 1. Januar 1900, und mein Vater
<lb/>war der Sohn eines orthodoxen russischen Priesters und meine
<lb/>Mutter eine Deutsche (ihren Eltern — und später ihrem Bru-
<lb/>der — gehörte ein großes Gut in Landskron am Fuße des
<lb/>Adlergebirges).</s></seg> <seg><s>Ich wurde zu Hause geboren,
damals machte
<lb/>man das noch so.</s></seg> <seg><s>Meine Hebamme war Magda, eine
Ungarin
<lb/>aus Preßburg.</s></seg> <seg><s>Meine Geburtsstadt
Brünn.</s></seg> <seg><s>Die genaue Stelle
<lb/>dann ein Bett, eine Pritsche im dritten Stock eines großen Miets-
<lb/>hauses in der Ferdinandstraße (später Masarykstraße, noch spä-
<lb/>ter Hermann-Göring-Straße und noch später wieder Masaryk-
<lb/>straße und noch später Straße des Sieges und dann kurz wieder
<lb/>Masarykstraße und dann wieder lange Zeit Straße des Sieges
<lb/>und heute wieder Masarykstraße).</s></seg></p>
<lb/><p><seg><s>Jede Geburt ist, falls Sie das nicht wissen sollten,
meine Lie-
<lb/>ben, für alle Beteiligten auch ein Zirkus von
Emotionen.</s></seg> <seg><s>Die He-
<lb/>bamme hielt mich und schrie etwas auf ungarisch und slo-
<lb/>wakisch.</s></seg> <seg><s>Mutter versuchte, mich anzusehen, und
obwohl man sie
<lb/>daran zu hindern versuchte und mit aller Gewalt ins Bett zurück-
<lb/>drückte, stützte sie sich auf die Ellenbogen, richtete sich auf
wie
<lb/>eine Kobra und murmelte einen Augenblick hastig etwas auf
<lb/>tschechisch und gleich darauf wieder auf deutsch.</s></seg>
<seg><s>Vater kniete auf
<lb/>dem Boden, und er, der es mit dem Glauben bisher nicht so genau
<lb/>genommen hatte, betete jetzt auf russisch, und in diesen großen
<lb/>orthodoxen Strom, der sich aus ihm herauswälzte, flossen auch
```

XSLT-Schablone

Angezeigt wird die XSLT-Schablone für die Transformation sowohl des tschechischen als auch des deutschen Kratochvil-Textes (Datei: `KR_nummerieren.xsl`). Die Zeilennummerierung wurde für die Darstellung in diesem Anhang zusätzlich hinzugefügt und gehört nicht zum eigentlichen Dateiinhalt.

```
1     <?xml version="1.0" encoding="UTF-8"?>
2     <xsl:stylesheet                                    version="1.0"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
xmlns="http://www.w3.org/TR/xhtml1/strict">
3     <xsl:output        method="xml"        omit-xml-declaration="yes"
standalone="yes"/>
4
5     <!-- Variable, mit der "Autoren-ID" -->
6     <!--  <xsl:variable name="autor">jku00de</xsl:variable> -->
7          <xsl:variable name="autor">jkn05cz</xsl:variable>
8
9     <!--Variable,  die  den  Anfangswert  für  den  Seitenzähler
      festlegt-->
10        <!-- der deutsche Text -->
11    <!-- <xsl:variable name="pagestart" select="6"/> -->
12        <!-- der tschechische Text -->
13         <xsl:variable name="pagestart" select="10"/>
14        <!-- ************ div(chapter)-Zähler  ************ -->
15        <xsl:template name="chapterzaehler">
16            <xsl:variable name="chapterzlr">
17                <xsl:number count="chapter"
18                    level="any"/>
19            </xsl:variable>
20            <xsl:value-of select="$chapterzlr"/>
21        </xsl:template>
22    <!-- ************ Absatz-Zähler ************ -->
23        <xsl:template name="pzaehler">
24            <xsl:variable name="pz">
25                <xsl:number count="p|lg|salute"
26                    level="any"/>
27            </xsl:variable>
28                <xsl:value-of select="$pz"/>
29        </xsl:template>
30    <!-- ************ Segment-Zähler  ************ -->
31        <xsl:template name="segzaehler">
32            <xsl:variable name="sgz">
33                <xsl:number count="seg"
34                    level="any"/>
35            </xsl:variable>
36            <xsl:value-of select="$sgz"/>
37        </xsl:template>
38    <!-- ************ Satz-Zähler  ************ -->
39        <xsl:template name="satzzaehler">
40            <xsl:variable name="satzz">
41                <xsl:number count="s"
42                    level="any"/>
43            </xsl:variable>
```

```
44            <xsl:value-of select="$satzz"/>
45        </xsl:template>
46    <!-- ************ Line-Zähler  ************ -->
47        <xsl:template name="linezaehler">
48            <xsl:variable name="linez">
49                <xsl:number count="l"
50                    level="any"/>
51            </xsl:variable>
52            <xsl:value-of select="$linez"/>
53        </xsl:template>
54    <!-- ************ Seiten-Zähler  ************ -->
55        <xsl:template name="seite_nr">
56            <xsl:variable name="pc">
57                <xsl:number count="pb" level="any"/>
58            </xsl:variable>
59            <xsl:value-of select="$pc + $pagestart"/>
60        </xsl:template>
61    <!-- ************ lb-Zähler  ************ -->
62        <xsl:template name="lbzaehler">
63            <xsl:variable name="lbzlr">
64                <xsl:number count="lb"
65                    from="pb"
66                    level="any"/>
67            </xsl:variable>
68            <xsl:value-of select="$lbzlr"/>
69        </xsl:template>
70    <!-- ********** Anfang der eigenen Schablone ********** -->
71        <xsl:template match="/">
72            <xsl:apply-templates/>
73        </xsl:template>
74        <xsl:template match="TEI">
75            <TEI>
76                <xsl:apply-templates/>
77            </TEI>
78        </xsl:template>
79        <xsl:template match="text">
80            <text>
81                <xsl:apply-templates/>
82            </text>
83        </xsl:template>
84        <xsl:template match="body">
85            <body>
86                <xsl:apply-templates/>
87            </body>
88        </xsl:template>
89        <xsl:template match="div">
90            <div>
91                <xsl:attribute name="type">
92                    <xsl:text>book</xsl:text>
93                </xsl:attribute>
94                <xsl:apply-templates/>
95            </div>
96        </xsl:template>
97        <xsl:template match="chapter">
98            <div>
99                <xsl:attribute name="type">
100                   <xsl:text>chapter</xsl:text>
```

```
101                 </xsl:attribute>
102                 <xsl:attribute name="xml:id">
103                     <xsl:text>div</xsl:text>
104                     <xsl:call-template name="chapterzaehler"/>
105                     <xsl:text>_</xsl:text>
106                     <xsl:value-of select="$autor"/>
107                 </xsl:attribute>
108                 <xsl:apply-templates/>
109             </div>
110         </xsl:template>
111         <xsl:template match="ersatz">
112             <div>
113                 <xsl:apply-templates/>
114             </div>
115         </xsl:template>
116         <xsl:template match="hi">
117             <hi>
118                 <xsl:attribute name="rend">
119                     <xsl:value-of select="@rend"/>
120                 </xsl:attribute>
121                 <xsl:apply-templates/>
122             </hi>
123         </xsl:template>
124         <xsl:template match="head">
125             <head>
126                 <xsl:apply-templates/>
127             </head>
128         </xsl:template>
129         <xsl:template match="pb">
130             <pb>
131                 <xsl:attribute name="n">
132                     <xsl:call-template name="seite_nr"/>
133                 </xsl:attribute>
134                 <xsl:apply-templates/>
135             </pb>
136         </xsl:template>
137         <xsl:template match="lb">
138             <lb>
139                 <xsl:attribute name="n">
140                     <xsl:call-template name="seite_nr"/>
141                     <xsl:text>:</xsl:text>
142                     <xsl:call-template name="lbzaehler"/>
143                 </xsl:attribute>
144                 <xsl:apply-templates/>
145             </lb>
146         </xsl:template>
147         <!--
148         dem "p"-Tag wird Attribut "xml:id" zugewiesen, und die
149         Varianten: <p xml:id="" rend="missing">   <p xml:id=""
150         rend=""> entstehen durch if-Abfrage
151         -->
152         <xsl:template match="p">
153             <p>
154                 <xsl:attribute name="xml:id">
155                     <xsl:text>div</xsl:text>
156                     <xsl:call-template name="chapterzaehler"/>
157                     <xsl:text>.p</xsl:text>
```

```
158                     <xsl:call-template name="pzaehler"/>
159                     <xsl:text>_</xsl:text>
160                     <xsl:value-of select="$autor"/>
161                 </xsl:attribute>
162                 <xsl:if test="@rend">
163                     <xsl:attribute name="rend">
164                         <xsl:value-of select="@rend"/>
165                     </xsl:attribute>
166                 </xsl:if>
167                 <xsl:apply-templates/>
168             </p>
169         </xsl:template>
170         <xsl:template match="lg">
171             <lg>
172                 <xsl:attribute name="xml:id">
173                     <xsl:text>div</xsl:text>
174                     <xsl:call-template name="chapterzaehler"/>
175                     <xsl:text>.lg</xsl:text>
176                     <xsl:call-template name="pzaehler"/>
177                     <xsl:text>_</xsl:text>
178                     <xsl:value-of select="$autor"/>
179                 </xsl:attribute>
180                 <xsl:apply-templates/>
181             </lg>
182         </xsl:template>
183         <xsl:template match="opener">
184             <opener>
185                 <xsl:apply-templates/>
186             </opener>
187         </xsl:template>
188         <xsl:template match="closer">
189             <closer>
190                 <xsl:apply-templates/>
191             </closer>
192         </xsl:template>
193         <xsl:template match="salute">
194             <salute>
195                 <xsl:attribute name="xml:id">
196                     <xsl:text>div</xsl:text>
197                     <xsl:call-template name="chapterzaehler"/>
198                     <xsl:text>.sal</xsl:text>
199                     <xsl:call-template name="pzaehler"/>
200                     <xsl:text>_</xsl:text>
201                     <xsl:value-of select="$autor"/>
202                 </xsl:attribute>
203                 <xsl:apply-templates/>
204             </salute>
205         </xsl:template>
206         <xsl:template match="seg">
207             <seg>
208                 <xsl:attribute name="xml:id">
209                     <xsl:text>div</xsl:text>
210                     <xsl:call-template name="chapterzaehler"/>
211                     <xsl:text>.p</xsl:text>
212                     <xsl:call-template name="pzaehler"/>
213                     <xsl:text>.seg</xsl:text>
214                     <xsl:call-template name="segzaehler"/>
```

```
215                     <xsl:text>_</xsl:text>
216                     <xsl:value-of select="$autor"/>
217                 </xsl:attribute>
218                 <xsl:if test="@rend">
219                     <xsl:attribute name="rend">
220                         <xsl:value-of select="@rend"/>
221                     </xsl:attribute>
222                 </xsl:if>
223                 <xsl:apply-templates/>
224             </seg>
225         </xsl:template>
226         <xsl:template match="s">
227             <s>
228                 <xsl:attribute name="xml:id">
229                     <xsl:text>div</xsl:text>
230                     <xsl:call-template name="chapterzaehler"/>
231                     <xsl:text>.p</xsl:text>
232                     <xsl:call-template name="pzaehler"/>
233                     <xsl:text>.seg</xsl:text>
234                     <xsl:call-template name="segzaehler"/>
235                     <xsl:text>.s</xsl:text>
236                     <xsl:call-template name="satzzaehler"/>
237                     <xsl:text>_</xsl:text>
238                     <xsl:value-of select="$autor"/>
239                 </xsl:attribute>
240                 <xsl:if test="@rend">
241                     <xsl:attribute name="rend">
242                         <xsl:value-of select="@rend"/>
243                     </xsl:attribute>
244                 </xsl:if>
245                 <xsl:apply-templates/>
246             </s>
247         </xsl:template>
248         <xsl:template match="l">
249             <l>
250                 <xsl:attribute name="xml:id">
251                     <xsl:text>div</xsl:text>
252                     <xsl:call-template name="chapterzaehler"/>
253                     <xsl:text>.lg</xsl:text>
254                     <xsl:call-template name="pzaehler"/>
255                     <xsl:text>.l</xsl:text>
256                     <xsl:call-template name="linezaehler"/>
257                     <xsl:text>_</xsl:text>
258                     <xsl:value-of select="$autor"/>
259                 </xsl:attribute>
260                 <xsl:apply-templates/>
261             </l>
262         </xsl:template>
263     </xsl:stylesheet>
```

Finale XML-Form

Die erste Seite der deutschen Fassung des Kratochvil-Textes, die Daten sind in der finalen XML-Form; es handelt sich um einen Ausschnitt aus der Datei: `zieldatei_KR_de_nummeriert.xml`

```
<div type="chapter" xml:id="div1_jku00de"><pb n="9"/><lb n="9:1"/>
<head>1
<lb n="9:2"/>Die Stimme meines Herrn</head>
<lb n="9:3"/><p xml:id="div1.p1_jku00de"><seg xml:id="div1.p1.seg1_
jku00de"><s xml:id="div1.p1.seg1.s1_jku00de">Geboren wurde ich, wenn
Sie das wirklich hören wol-
<lb n="9:4"/>len, meine Herren, geboren wurde ich in der Nacht vom
<lb n="9:5"/>31. Dezember 1899 auf den 1. Januar 1900, und mein Vater
<lb n="9:6"/>war der Sohn eines orthodoxen russischen Priesters und
meine
<lb n="9:7"/>Mutter eine Deutsche (ihren Eltern — und später ihrem
Bru-
<lb n="9:8"/>der — gehörte ein großes Gut in Landskron am Fuße des
<lb n="9:9"/>Adlergebirges).</s></seg> <seg xml:id="div1.p1.seg2_
jku00de"><s xml:id="div1.p1.seg2.s2_jku00de">Ich wurde zu Hause
geboren, damals machte
<lb n="9:10"/>man das noch so.</s></seg> <seg xml:id="div1.p1.seg3_
jku00de"><s xml:id="div1.p1.seg3.s3_jku00de">Meine Hebamme war Magda,
eine Ungarin
<lb n="9:11"/>aus Preßburg.</s></seg> <seg xml:id="div1.p1.seg4_
jku00de"><s xml:id="div1.p1.seg4.s4_jku00de">Meine Geburtsstadt
Brünn.</s></seg> <seg xml:id="div1.p1.seg5_jku00de"><s
xml:id="div1.p1.seg5.s5_jku00de">Die genaue Stelle
<lb n="9:12"/>dann ein Bett, eine Pritsche im dritten Stock eines
großen Miets-
<lb n="9:13"/>hauses in der Ferdinandstraße (später Masarykstraße,
noch spä-
<lb n="9:14"/>ter Hermann-Göring-Straße und noch später wieder
Masaryk-
<lb n="9:15"/>straße und noch später Straße des Sieges und dann kurz
wieder
<lb n="9:16"/>Masarykstraße und dann wieder lange Zeit Straße des
Sieges
<lb n="9:17"/>und heute wieder Masarykstraße).</s></seg></p>
<lb n="9:18"/><p xml:id="div1.p2_jku00de"><seg xml:id="div1.p2.seg6_
jku00de"><s xml:id="div1.p2.seg6.s6_jku00de">Jede Geburt ist, falls
Sie das nicht wissen sollten, meine Lie-
<lb n="9:19"/>ben, für alle Beteiligten auch ein Zirkus von
Emotionen.</s></seg> <seg xml:id="div1.p2.seg7_jku00de"><s xml:id="
div1.p2.seg7.s7_jku00de">Die He-
<lb n="9:20"/>bamme hielt mich und schrie etwas auf ungarisch und slo-
<lb n="9:21"/>wakisch.</s></seg> <seg xml:id="div1.p2.seg8_jku00de"><s
xml:id="div1.p2.seg8.s8_jku00de">Mutter versuchte, mich anzusehen, und
obwohl man sie
<lb n="9:22"/>daran zu hindern versuchte und mit aller Gewalt ins Bett
zurück-
<lb n="9:23"/>drückte, stützte sie sich auf die Ellenbogen, richtete
sich auf wie
<lb n="9:24"/>eine Kobra und murmelte einen Augenblick hastig etwas
auf
```

```
<lb n="9:25"/>tschechisch und gleich darauf wieder auf
deutsch.</s></seg> <seg xml:id="div1.p2.seg9_jku00de"><s xml:id="div1.
p2.seg9.s9_jku00de">Vater kniete auf
<lb n="9:26"/>dem Boden, und er, der es mit dem Glauben bisher nicht
so genau
<lb n="9:27"/>genommen hatte, betete jetzt auf russisch, und in diesen
großen
<lb n="9:28"/>orthodoxen Strom, der sich aus ihm herauswälzte, flossen
auch
```

TEI-Header

Zur Illustration wird hier nur der Header zur deutschen Datei des Romans von Kratochvil angeführt. Der Header wird in der Datei `header_krat_de.xml` gespeichert und erst am Ende der XML-Bearbeitung mit dem übrigen Roman-Text verbunden. Die Zeilennummerierung wurde für die Darstellung in diesem Anhang zusätzlich hinzugefügt und gehört nicht zum eigentlichen Dateiinhalt.

```
1     <?xml version="1.0" encoding="UTF-8"?>
2     <?oxygen RNGSchema="myTEI.rnc" type="compact"?>
3     <TEI xmlns="http://www.tei-c.org/ns/1.0">
4         <teiHeader type="text">
5             <fileDesc>
6                 <titleStmt>
7                     <title>Unsterbliche Geschichte oder Das Leben der
      Sonja Trotzkij-Sammler oder Karneval</title>
8                     <author>Jiří Kratochvil</author>
9                     <editor role="translator">Kathrin Liedtke</editor>
10                                    <editor  role="translator">Milka
      Vagadayová</editor>
11                    <respStmt>
12                        <resp>Digitalisierung des Textes</resp>
13                        <name>Josef Molnár</name>
14                    </respStmt>
15                    <respStmt>
16                        <resp>Korrekturlesen</resp>
17                        <name>Veronika Kotůlková</name>
18                    </respStmt>
19                </titleStmt>
20                <publicationStmt>
21                    <publisher>Ammann Verlag</publisher>
22                    <pubPlace>Zürich</pubPlace>
23                    <date>2000</date>
24                </publicationStmt>
25                <sourceDesc>
26                      <bibl>Kratochvil, Jiří: Unsterbliche Geschichte
      oder Das Leben der Sonja Trotzkij-Sammler oder Karneval. Zurüch
      2000, S. 7 - 296.</bibl>
27                    <biblFull>
28                        <titleStmt>
29                              <title>Unsterbliche Geschichte oder Das
      Leben der Sonja Trotzkij-Sammler oder Karneval</title>
30                            <author>Jiří Kratochvil</author>
31                        </titleStmt>
32                        <publicationStmt>
33                             <publisher xml:id="Rechteinhaber">Ammann
      Verlag</publisher>
34                            <pubPlace>Zürich</pubPlace>
35                            <date>2000</date>
36                        </publicationStmt>
37                    </biblFull>
38                </sourceDesc>
39            </fileDesc>
40            <encodingDesc>
```

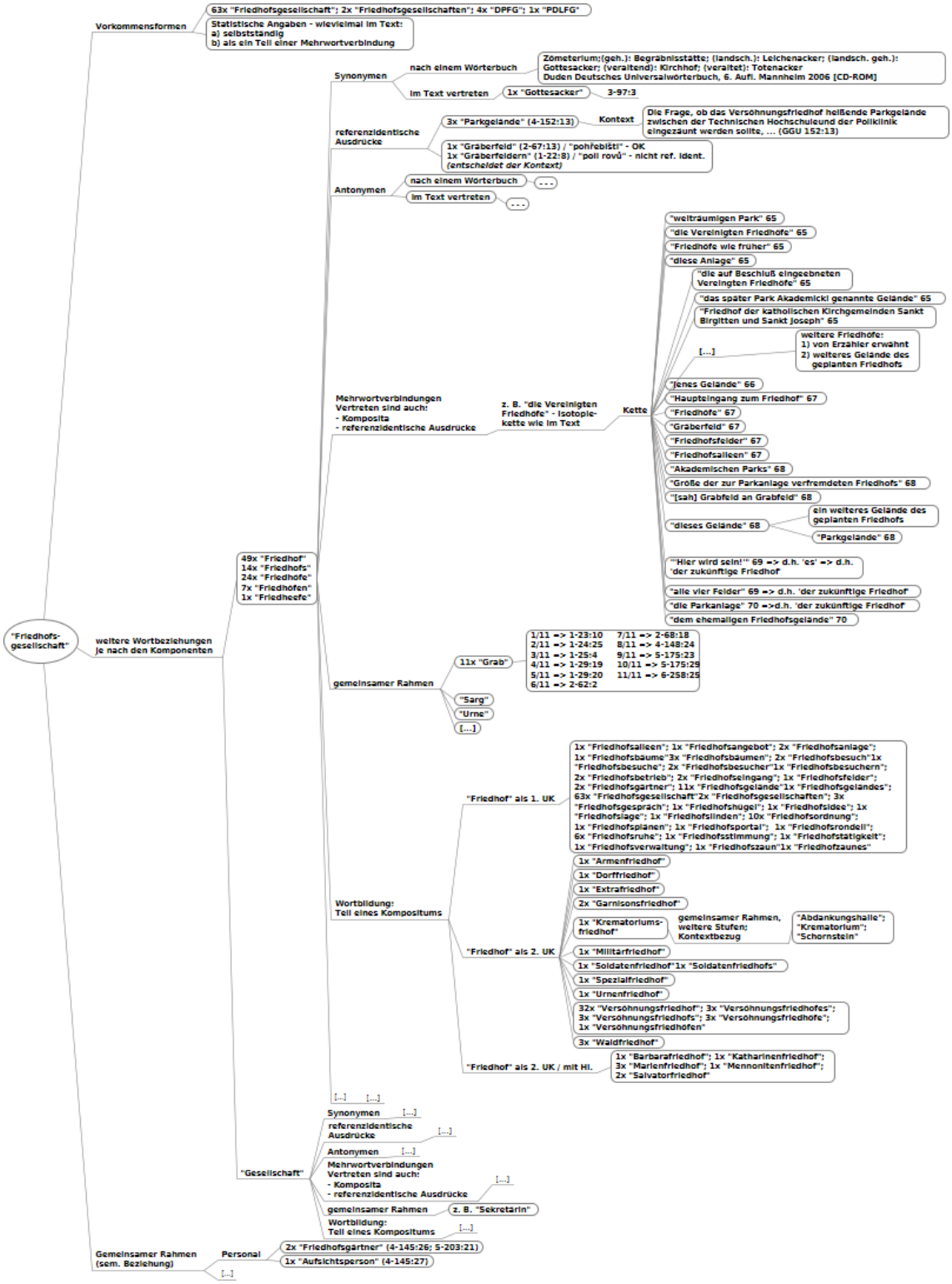
```
41              <projectDesc>
42                   <p>Der bearbeitete Text dient als Grundlage für
     das DeuCze-Korpus.</p>
43              </projectDesc>
44              <samplingDecl>
45                   <p>Der Gesamttext des Romans wurde gescannt und
     bearbeitet mit einem OCR-Programm. Die gewonnene Textdaten wurden
     mit einem PHP-Skript in eine XML-Datei umgewandelt, die als
     Ausgangsbasis für die TEI P5-Datei dient.</p>
46              </samplingDecl>
47              <editorialDecl>
48                  <normalization>
49                       <p>Gestaltung des Originaltextes, die in der
     Textkodierung nicht kodiert wird:
50                       - die Bücher- und Kapitelüberschriften sind in
     Kapitälchenschrift gesetzt, zentriert
51                       - jeder erste Absatz in dem jeweiligen Kapitel
     beginnt mit einer Initiale
52                       - jeder weitere Absatz ist links eingezogen
53                        - die Verse sind zentriert (z. B. auf S. 68,
     172, 178, 188)
54                       - der Abschiedsgruß in dem Brief (S. 142) ist
     rechtsbündig
55                        - die Zahlen bei der Seitennummerierung sind
     zentriert.
56                                    - Die Tierbezeichnungen in den
     Buchüberschriften sind auch im Originaltext mit Großbuchstaben,
     nicht mit Kapitälchenschrift, gesetzt.
57                       - Die Leerseiten (S. 8, 70, 128, 190 und 248)
     sind nur durch das Pagebreak-Tag repräsentiert</p>
58                  </normalization>
59                  <segmentation>
60                      <p>
61                        In dem Text werden a) hierarchische und b)
     lineare Strukturen des Quelltextes kodiert. Für die
     Beschreibungszwecke innerhalb des Headers stehen die Tags in
     runden Klammern
62                       </p><p>
63                       a) hierarchische Strukturen:
64                       - Romanbücher: (div type="book")(/div)
65                       - Kapitel: (div type="chapter")(/div)
66                       - Kapitelüberschriften: (head)(/head)
67                       - Absätze: (p)(/p)
68                       - Segmente: (seg)(/seg)
69                       - Sätze: (s)(/s)
70                       - Gedichte: (lg)(/lg)
71                       - Verszeilen: (l)(/l)
72                       - Brief: als ein Ganzes in (div)(/div)
73                            - Anfang-Grußformel des Briefes: (opener)
     (salute)(/salute)(/opener)
74                             - Ende-Grußformel des Briefes: (closer)
     (salute)(/salute)(/closer)
75                       </p><p>
76                       b) lineare Strukturen:
77                       - Seitenumbruch: (pb/)
78                       - Zeilenumbruch: (lb/)
79                        - Formatierung der Kursiv- (hi rend="italic")
     (/hi) und Kapitälchenschrift (hi rend="smallCaps")(/hi)
80                       </p>
```

```
81                    </segmentation>
82                </editorialDecl>
83            </encodingDesc>
84            <profileDesc>
85                <langUsage>
86                    <language ident="de-DE">Deutsch</language>
87                </langUsage>
88            </profileDesc>
89            <revisionDesc>
90                              <change  xml:id="S.103"  when="20090708"
      who="#JM">Korrektur   eines   vermutlichen   Satzfehlers   im
      Originaltext:
91                    - ursprünglich:
92                    "sei mit Zügen, Bahnhöfen und Bahnsteigen geredezu
      gespickt."
93                    - korrigiert:
94                    "sei mit Zügen, Bahnhöfen und Bahnsteigen geradezu
      gespickt."
95                </change>
96                              <change  xml:id="S.250"  when="20090708"
      who="#JM">Korrektur   eines   vermutlichen   Satzfehlers   im
      Originaltext:
97                    - ursprünglich:
98                    "dern vom Beginn des Jahrhunders, mit dem Rucksack
      auf dem"
99                    - korrigiert:
100                         "dern vom Beginn des Jahrhunderts, mit dem
Rucksack auf dem"
101               </change>
102                             <change  xml:id="S.261"  when="20090708"
who="#JM">Korrektur eines vermutlichen Satzfehlers im Originaltext:
103                   - ursprünglich:
104         "tin, wenn ich dir alles erzählen sollte, was ich in in den"
105                   - korrigiert:
106         "tin, wenn ich dir alles erzählen sollte, was ich in den"
107               </change>
108                             <change  xml:id="S.277"  when="20090708"
who="#JM">Korrektur eines vermutlichen Satzfehlers im Originaltext:
109                   - ursprünglich:
110         "dert und als MUnd waschinenführer dann alle möglichen"
111                   - korrigiert:
112         "dert und als Maschinenführer dann alle möglichen"
113               </change>
114                             <change  xml:id="S.279"  when="20090708"
who="#JM">Korrektur eines vermutlichen Satzfehlers im Originaltext:
115                   - ursprünglich:
116         "schen dem russisch-orthodoxon liturgischen Pomp (einer Ek-"
117                   - korrigiert:
118         "schen dem russisch-orthodoxen liturgischen Pomp (einer Ek-"
119               </change>
120           </revisionDesc>
121       </teiHeader>
122       <text><body>
123    [Text des Romans]
124       </body></text>
125    </TEI>
```

Wortbildungsaktivität zu *Friedhof*

Die Zahlen vor den Belegen geben die Anzahl der Tokens, also die Vorkommensfrequenz, im Gesamttext, an.

Wortschatz zu *Friedhofsgesellschaft*



Lebenslauf

Persönliche Daten

Name	Josef Molnár
Geburtstag	6. Januar, 1981
Geburtsort	Brno, Tschechische Republik
Anschrift	Dvorní 10 Lužice CZ-696 18
E-Mail	josefmlnr@gmail.com

Studium

2007–2010	Schlesische Universität in Opava Promotionsstudium: Korpuslinguistik – Deutsch
2002–2007	Schlesische Universität in Opava Germanistik

Schulbildung

1995–2000	Hotelfachschule – Fremdenverkehrmanagement

Andere Daten

SS 2005	Austauschstudent an der Julius-Maximilians-Universität Würzburg; (im Rahmen des ERASMUS-Programms)
ak. Jahr 2008/09	Julius-Maximilians-Universität Würzburg Studienaufenthalt im Rahmen des Austauschprogramms ‚Studienbörse Germanistik‘